Tânia Schmidt

Planejamento de Capacidade em Provedores de Serviços Internet

Florianópolis – SC
2000

UNIVERSIDADE FEDERAL DE SANTA CATARINA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Tânia Schmidt

Planejamento de Capacidade em Provedores de Serviços Internet

Dissertação submetida à Universidade Federal de Santa Catarina como parte dos requisitos para a obtenção do grau de Mestre em Ciência da Computação.

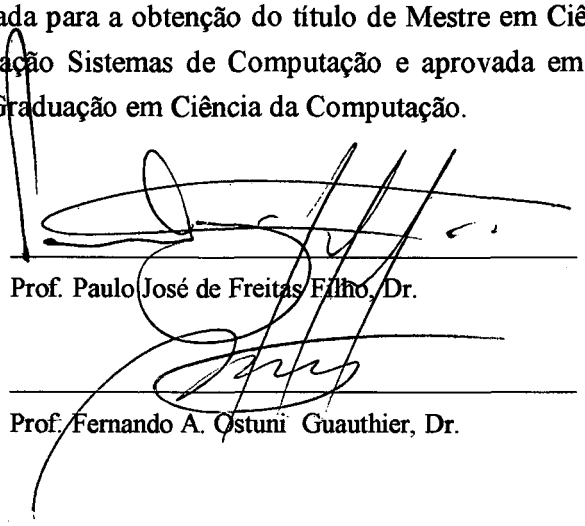
Prof. Dr. Paulo José de Freitas Filho

Florianópolis, setembro 2000.

Planejamento de Capacidade em Provedores de Serviços Internet

Tânia Schmidt

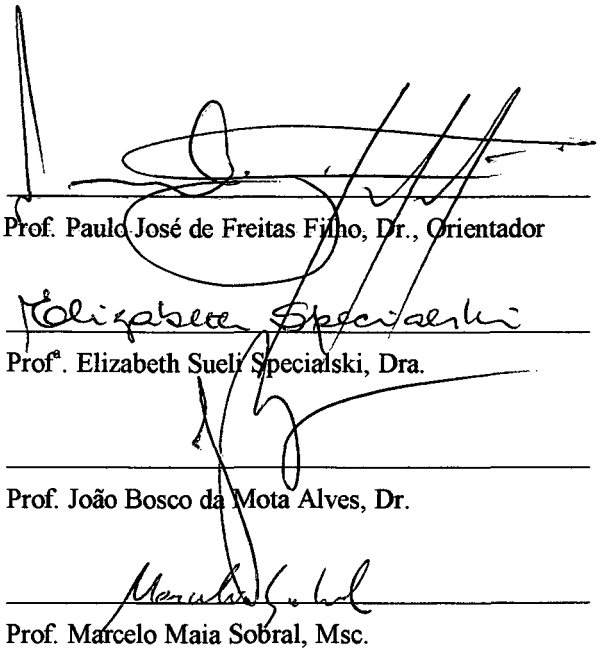
Esta Dissertação foi julgada adequada para a obtenção do título de Mestre em Ciência da Computação Área de Concentração Sistemas de Computação e aprovada em sua forma final pelo Programa de Pós-Graduação em Ciência da Computação.



Prof. Paulo José de Freitas Filho, Dr.

Prof. Fernando A. Ostuni Guauthier, Dr.

Banca Examinadora:



Prof. Paulo José de Freitas Filho, Dr., Orientador

Profª. Elizabeth Sueli Specialski, Dra.

Prof. João Bosco da Mota Alves, Dr.

Prof. Marcelo Maia Sobral, Msc.

"Mesmo uma grande jornada se inicia com um primeiro passo."

Provérbio chinês

Dedico este trabalho

Aos meus maravilhosos pais,

Lídia Maria Schmidt e

Velácio Luiz Schmidt

Agradecimentos

À Deus, pela vida.

Aos meus pais, por todo o amor, segurança, confiança e sobretudo pelo apoio em todas as minhas decisões.

As minhas irmãs, Nádia e Susi, por todo carinho, paciência, apoio e confiança.

Ao meu orientador, professor Paulo José de Freitas Filho, pelo alto astral, paciência, colaboração e incentivo na realização do trabalho.

Aos professores Elizabeth Sueli Specialski, João Bosco da Mota Alves e Marcelo Maia Sobral pela participação como membros da banca.

À Graziela Naspolini Delpizzo e Marcelo Maia Sobral pelas contribuições durante a realização do trabalho.

À Arthur Fernando Dellagiustina Lago, proprietário da empresa Intercorp – Internet Provider, pela disponibilidade e diversas informações sobre a empresa, permitindo a realização deste trabalho.

Aos amigos e colegas do Banco do Brasil que sempre acreditaram em mim, pelo incentivo e pelas horas que tive que me ausentar.

Às minhas amigas do coração, Cândida, Enis, Gilmara, Lili, Luciane, Rosi e Sibele (em ordem alfabética para ninguém ficar com ciúmes, por favor) por toda a amizade, companheirismo, palavras de incentivo, confiança e pela alegria de viver.

Aos demais amigos e amigas, pelos ótimos momentos, pelas viagens, churrascadas, festas e pela companhia quando a MH aparece.

A todas as pessoas que passaram pela minha vida e que me ensinaram alguma coisa. Afinal, como diz aquele velho dizer: "Cada pessoa que passa pela nossa vida não nos deixa sós, deixa um pouco de si, levando um pouquinho da gente".

Sumário

Lista de Tabelas

Lista de Figuras

Resumo

O planejamento de capacidade é o processo de prever quando o nível de uma carga futura irá saturar o sistema e determinar o melhor custo x benefício de evitar esta saturação. Tal previsão é baseada no processo de evolução natural da carga de trabalho do sistema existente, desenvolvimento de aplicações/serviços e trocas no comportamento dos consumidores em função da disponibilidade de novas funções de negócios.

Esta dissertação apresenta uma aplicação de tal processo no ambiente de provimento de serviços Internet com o objetivo de mostrar a importância da utilização da metodologia no auxílio às organizações, possibilitando tomada de decisões mais seguras, baseadas em avaliações técnicas de seus ambientes. A escolha do ambiente foi feita pela importância dos provedores de serviço Internet na qualidade dos serviços oferecidos na web. Isto é justificável, pois, dependendo do tráfego e capacidade instalada no provedor, problemas de desempenho podem vir a ocorrer, contribuindo para o incremento do tempo de resposta aos usuários finais. Os principais resultados do trabalho são: uma proposta de adaptação da metodologia e a sua respectiva aplicação com a descrição detalhada de cada uma das etapas, permitindo a utilização em outros ambientes.

Abstract

The capacity planning is a process that predict when the level of a future workload will overwhelm the system and determine the ideal cost x benefit in order to avoid this saturation. Such prediction is based on the process of natural evolution of workload of the existing system, on the development of applications/services and on changes of consumers' behavior according to the availability of new business functions.

This written essay presents an application of such process in the environment of Internet services provision with the purpose of indicating the importance of the utilization of methodology to back companies up, enabling the reach of more certain decisions, based on technical appraisal of their spheres. The choice of the environment was made due to the importance of Internet services providers on quality of services offered in web. This is justifiable, because depending on the transit and on the provider's installed capacity, performance problems can occur, contributing to the increase of the answer time to final users. The main results of the essay are: an adaptation proposal of the methodology and the detailed description of each and every step, allowing its applications in other environments.

CAPÍTULO 1

Introdução

Num primeiro momento, a expressão planejamento de capacidade parece um termo relativamente novo no mundo da computação. No entanto é uma metodologia que vem sendo praticada na informática desde a época do *mainframe*. Tendo ficado relativamente esquecida por um período, observa-se hoje a sua retomada, considerando as necessidades de adequação entre a demanda crescente de serviços e a capacidade de provê-los.

Há poucos anos a maioria das empresas trabalhava de maneira completamente reativa. Os *upgrades* eram feitos de maneira aproximada, fazendo com que os sistemas inclusive ficassem fora de uso por certos períodos. Nos dias atuais isso não é mais aceitável pois a maioria das aplicações são de missão crítica e a complexidade dos sistemas aumentou muito. Hoje têm-se diferentes tipos de computadores, com sistemas operacionais próprios, comunicando-se através dos vários tipos de redes, utilizando uma variedade de técnicas de comunicação. Não é simples adicionar um novo *hardware*, *software*, prever um aumento da carga de trabalho e mudança no comportamento dos usuários neste contexto.

Este conjunto de fatores fez surgir nas organizações a necessidade de gerenciar estas trocas e mudanças em seus ambientes de uma maneira mais organizada O presente estudo busca exatamente mostrar como a utilização da metodologia de planejamento de capacidade e avaliação de desempenho pode auxiliar as organizações na tomada de decisões mais seguras, baseadas em avaliações técnicas de seus ambientes, permitindo fornecer desta forma, serviços de qualidade aos seus usuários. Uma das áreas em que esta metodologia vem sendo cada vez mais aplicada é o ambiente de serviços Internet, objeto deste trabalho.

1.1 Problema

O problema a ser aqui tratado envolve o encaminhamento de soluções viáveis sob o ponto de vista do desempenho com a aplicação de técnicas e métodos de planejamento de capacidade e avaliação de desempenho, a provedores de serviços Internet.

1.2 Justificativa

Os serviços Internet estão crescendo de maneira exponencial em todo o mundo, principalmente após a aposta no comércio eletrônico. Os números que apontam o potencial de realização de negócios pela Internet são surpreendentes e não páram de crescer. Com toda esta expectativa as empresas e instituições inseridas no mundo virtual não podem mais se contentar em apenas estar presente na realidade "Internet". Não importa se a empresa trabalha com *e-commerce*, *e-business*, desenvolve *web sites* ou é uma provedora de serviços. Faz-se necessário cada vez mais sustentar esta presença, fornecer serviços de qualidade e planejar constantemente o seu crescimento, antevendo as necessidades dos usuários.

Para não serem surpreendidas pela demanda as empresas começaram a investir cada vez mais em infra-estrutura e gerenciamento, passando a se preocupar cada vez mais com desempenho. No entanto ter o último e mais rápido equipamento nem sempre garante o melhor desempenho e isto em muitos casos também não é economicamente viável. Como alternativa as organizações podem obter incremento de desempenho selecionando a correta combinação de equipamentos e cuidadosa otimização dos recursos disponíveis.

Além disso, esta-se numa época em que o tempo é fator cada vez mais preponderante na eficácia das empresas e ferramentas, métodos e planejamentos, enfim, que o otimizem são cada vez mais requisitos indispensáveis.

Para auxiliar no gerenciamento deste crescimento e no provimento destes novos serviços e necessidades, justifica-se este trabalho ressaltando a importância da realização do processo de planejamento de capacidade e exemplificando-o através de um estudo de caso.

1.3 Objetivos

1.3.1 Objetivo geral

Esta dissertação tem como objetivo a aplicação da metodologia de planejamento de capacidade e avaliação de desempenho no ambiente de provimento de serviços Internet.

1.3.2 Objetivos específicos

- Propor uma adequação da metodologia de planejamento de capacidade;
- Estudar as diferentes tecnologias de transmissão, os serviços envolvidos no provimento Internet, a arquitetura e a topologia do ambiente Internet;
- Possibilitar um planejamento pró-ativo do sistema;
- Estimar se o ambiente deverá, ou não, mudar e até onde e quanto poderá ser expandido;
- Desenvolver modelos de desempenho e de carga no planejamento de capacidade destes ambientes.

1.4 Abrangência

As maiores contribuições deste trabalho são:

- Mostrar a importância da realização do processo de planejamento de capacidade;

- Apresentar uma proposta de adaptação da metodologia com a respectiva descrição das etapas aplicadas num ambiente real, permitindo a sua reprodução em outros ambientes.

Entretanto, foram necessárias realizar algumas simplificações:

- Somente foi disponibilizado um *log* inexpressivo da carga real do ambiente, tendo sido necessário utilizar dados de um ambiente similar e complementá-los com informações colhidas na literatura;

- A ferramenta utilizada para modelagem é uma versão acadêmica e apresenta limitações quanto a quantidade de recursos que podem ser utilizados no modelo;

- No modelo de desempenho foram realizadas simplificações. A política de substituição de arquivos do servidor *proxy* não é modelada;

- Não é feita uma previsão da carga futura baseada em tendência passada, pelo fato de não existir em *log* com a carga passada;

- O modelo de custos, os planos de configuração, pessoal e investimentos previstos na metodologia não são realizados.

1.5 Estrutura da dissertação

Este documento está estruturado em seis capítulos: Introdução, Planejamento de Capacidade e Avaliação de Desempenho de Sistemas, Ambiente de Provimento de Serviços Internet, Estudo de Caso, Experimentação e Análise dos Resultados e por último, a Conclusão.

O capítulo 1 – Introdução – realiza uma introdução geral, apresentando o problema que motivou a realização da dissertação, a justificativa, os objetivos, a abrangência e a sua estrutura.

O capítulo 2 - Planejamento de Capacidade e Avaliação de Desempenho de Sistemas – apresenta a contextualização, objetivos e metodologia de planejamento de capacidade e avaliação de desempenho de sistemas e ao final, uma proposta de adequação da metodologia de Planejamento de Capacidade hoje utilizada.

O capítulo 3 – Ambiente de Provimento de Serviços Internet – descreve a arquitetura Internet e suas quatro camadas, contextualiza um provedor de serviços, seus componentes e meios de conexão.

O capítulo 4 – Estudo de Caso – relata a aplicação do processo num ambiente de provimento de serviços Internet, descrevendo as etapas de compreensão do ambiente, definição dos objetivos dos negócios, caracterização da carga de trabalho e por fim, uma previsão futura da carga.

O capítulo 5 – Experimentação e Análise dos Resultados – descreve as demais etapas como a modelagem do ambiente, definição dos fatores, níveis, métricas e projeto experimental utilizado. Em seguida é realizada uma análise dos resultados e previsão do comportamento futuro do ambiente.

O capítulo 6 – Conclusão – são apresentadas as conclusões da dissertação, trazidas as dificuldades e sugeridas propostas para estudos futuros.

CAPÍTULO 2

PLANEJAMENTO DE CAPACIDADE E AVALIAÇÃO DE DESEMPENHO DE SISTEMAS

Este capítulo tem por objetivo explorar as metodologias de planejamento de capacidade e avaliação de desempenho. Para tanto é dividido em três seções principais. A primeira delas trata sobre planejamento de capacidade, iniciando com a conceituação, apresentação dos objetivos, descrição da metodologia e, ao final, as dificuldades usualmente encontradas neste processo. A segunda seção conceitua o processo de avaliação de desempenho, seus objetivos e descreve a metodologia.

A última seção é dedicada à apresentação de uma proposta de ajuste da metodologia de planejamento de capacidade com a incorporação de alguns passos do processo de avaliação de desempenho. Ao final são apresentadas as justificavas que motivaram a apresentação da alteração na metodologia original.

2.1 Planejamento de capacidade

No passado, de acordo com SALSBURG (1997) muitos administradores de sistemas simplesmente reagiam aos problemas de desempenho de seus computadores, quando alguma coisa errada acontecia. Era necessário analisar os dados históricos para encontrar a causa da degradação do desempenho. O planejamento de capacidade surgiu como uma alternativa para realizar um gerenciamento pró-ativo, identificando os possíveis problemas antes que os mesmos viessem a ocorrer, permitindo desta maneira evitá-los.

Segundo MENASCÉ & ALMEIDA (1998) planejamento de capacidade "é o processo que visa prever se e quando o nível de carregamento futuro do sistema estará saturado, considerando os aspectos de custo/benefício e o tempo que levará para o sistema saturar". Esta previsão é baseada no processo de evolução natural da carga de trabalho, desenvolvimento de novas aplicações/serviços e mudanças no comportamento dos consumidores. Sendo assim, pode-se dizer que o planejamento de capacidade ajuda a prever situações sobre o comportamento dos sistemas em determinadas situações, como, por exemplo, o aumento da carga de trabalho.

Planejamento requer decisões, a atividade de decidir requer necessariamente várias habilidades para analisar componentes e situações, sintetizando-as em um todo. Para isso, BROWNING (1995) diz que, além da dimensão científica envolvida no processo de planejamento de ambientes de sistemas computacionais, existe também uma dimensão artística, ou seja, além de considerar os fatos, deve-se utilizar a perspicácia. Segundo o autor, a ciência do planejamento de capacidade envolve várias técnicas disponíveis na informática e na estatística, bem como a utilização de uma metodologia a ser seguida. A dimensão "objetiva" do planejamento de capacidade pode incluir análises matemáticas complexas, envolvendo técnicas de previsão, análises estatísticas, ou causais, e análise de informações com o objetivo de extrair modelos e tendências. A "arte" do planejamento requer uma compreensão mais subjetiva dos acontecimentos, dos fatores humanos envolvidos e uma habilidade para colocar em prática os planos realizados e partir para uma ação efetiva. Envolve, por exemplo, estimar a probabilidade de que alguns projetos podem ser aprovados e que outros podem falhar; prever antecipadamente o crescimento ou declínio em várias atividades dos negócios que afetam a carga de trabalho.

JAIN (1991) ainda complementa, afirmando que um planejamento de capacidade eficaz requer o entendimento do relacionamento, às vezes conflitante, entre as necessidades do negócio, a carga computacional, a capacidade computacional e o nível de serviço requerido.

2.1.1 Objetivos do planejamento de capacidade

O planejamento de capacidade visa principalmente gerar um nível aceitável de serviço computacional à organização, ao responder às demandas de carga geradas pelo sistema, permitindo melhorias neste com a adequação de *hardware* e *software*, com vistas a evitar a deficiência de capacidade.

Uma importante consideração sobre a razão de realizar o planejamento de capacidade é colocada por DOMANSKI (1999): "Se um sistema começa a ficar sobrecarregado e o desempenho torna-se lento, os usuários ficarão insatisfeitos e farão a opção por um concorrente que ofereça melhores condições de serviço". Isto por si só já bastaria para que as organizações se preocupassem muito mais com a capacidade de desempenho de seus sistemas. Segundo LOPES (2000), no ambiente web isto é ainda mais crítico, pois um cliente chega ao concorrente com apenas um clique no mouse e garantir alta disponibilidade vinte e quatro horas durante, sete dias por semana, é essencial.

Um outro motivo para planejar com antecedência é que a solução de problemas de desempenho não ocorre instantaneamente. Mesmo que a empresa possua recursos financeiros suficientes para comprar novos equipamentos e *softwares*, se um planejamento não for realizado, os problemas irão aparecer novamente em pouco tempo.

WALDNER (1997) ainda completa afirmando que "um bom planejamento de capacidade tem por objetivo trazer um retorno financeiro positivo sobre os investimentos realizados pela empresa". Com a realização de um planejamento prévio, a empresa pode reduzir os custos de operação na compra de *hardware* e *software* adequados às suas funções, evitando gastos desnecessários; não terá queda de produtividade para os usuários finais nas suas funções críticas, garantindo a qualidade dos serviços prestados. Estes fatores tornam a empresa mais competitiva, preserva a sua imagem externa e possibilita um aumento de sua participação no mercado.

Segundo BROWNING (1995), os objetivos do planejamento de capacidade estão completamente inter-relacionados com o planejamento de negócios e a previsão de atividades futuras de uma organização, pois estes tem estreita relação com a demanda de recursos computacionais. Usando modelos estatísticos, estimativas do volume de negócios futuros podem ser diretamente transpostos para quantidades específicas de recursos computacionais. Por este motivo os responsáveis pela realização do planejamento de capacidade devem utilizar em seus modelos um ou mais elementos de negócios como uma variável independente.

2.1.2 Metodologia de planejamento de capacidade

Com o *hardware*, *software* e linhas de comunicação cada vez mais baratos e de fácil acesso, o planejamento de uma rede parece ser muito simples. O uso de ferramentas para planejar o crescimento pode parecer desnecessário. Entretanto, com o surgimento da Internet, a demanda cada vez maior por serviços de qualidade e a complexidade dos ambientes distribuídos têm levado as organizações a prestar mais atenção a esta metodologia.

Para um planejamento de capacidade adequado, MENASCÉ & ALMEIDA (1998) sugerem a adoção da seguinte metodologia:

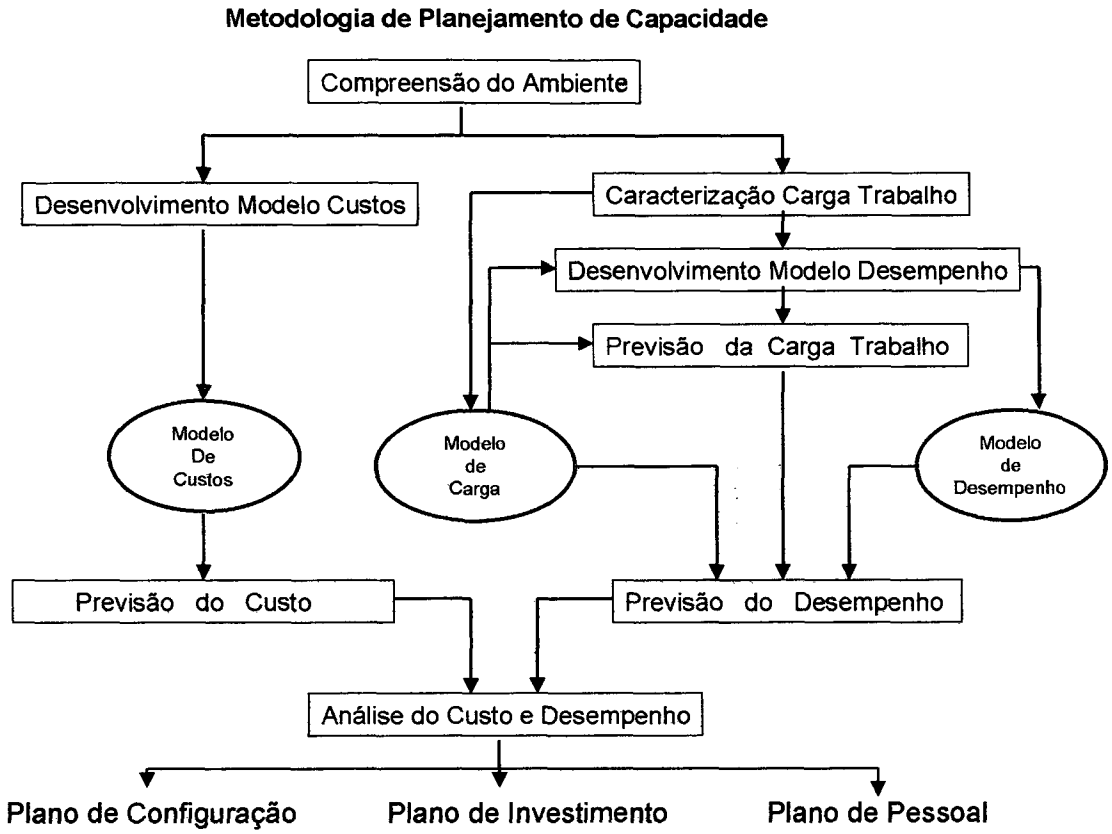


FIGURA 1 – Metodologia de Planejamento de Capacidade

2.1.2.1 Compreensão do ambiente

Algumas instalações computacionais são gerenciadas de modo completamente reativo. Nenhum problema é previsto, planejado ou corrigido até se tornar uma crise. Um gerenciamento bem sucedido de recursos computacionais deve ser pró-ativo, ou seja, toma-se uma abordagem organizada e planejada para cada esforço no sentido de evitar crises. Para uma gerência pró-ativa, deve-se entender as necessidades atuais da organização, compreender a carga de trabalho, o desempenho do sistema atual processando esta carga e as expectativas de serviços dos usuários. Isto consiste no estudo do *hardware* (clientes e servidores), do *software* (sistemas operacionais, *middleware* e aplicações), dos elementos de conectividade e dos protocolos envolvidos no ambiente. Com relação ao tráfego, os períodos de pico, as estruturas de gerência e os

níveis de qualidade de serviços definidos. Em resumo, deve-se entender a situação atual do sistema antes de poder planejar o futuro.

A obtenção destas informações muitas vezes não é uma tarefa fácil. Reuniões, entrevistas, questionários, *logs*, documentos de planejamento e projetos podem e dever ser utilizados na busca de informações.

2.1.2.2 Caracterização da carga de trabalho

É a descrição da lista de todas as entradas a que o sistema está submetido, considerando seus principais componentes, com a seleção daqueles que melhor identificam a carga em relação à análise. Algumas questões como serviços praticados pelo sistema, nível de detalhamento, componentes básicos, representatividade e atualidade são examinadas.

2.1.2.2.1 Passos gerais para a caracterização da carga

Embora cada sistema possa requerer um conjunto específico de características na caracterização da carga, existem algumas linhas gerais que podem ser bem aplicadas em todos os tipos de sistemas. O processo de criação do modelo indica que o resultado de um determinado estágio pode indicar a necessidade de revisar um estágio anterior (*vide fig. 2*). O diagrama indica um possível fluxo:

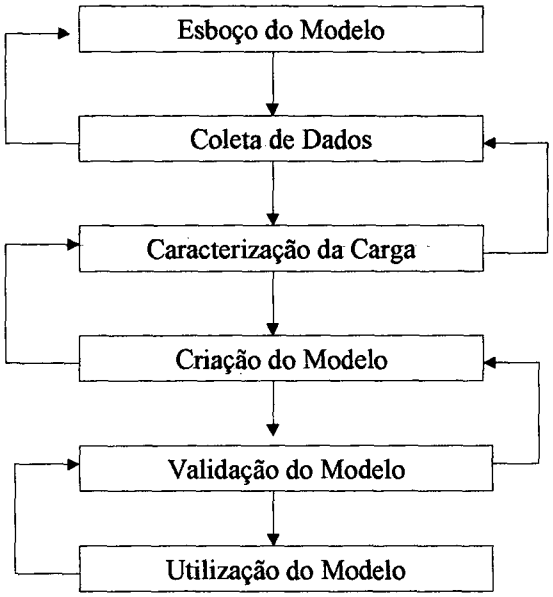


FIGURA 2 – Passos para caracterização da carga de trabalho

No primeiro passo, "Esboço do Modelo", um esboço da idéia inicial do sistema é formulado baseado no propósito do estudo. Algumas questões como sessão de medição, nível de modelagem, componentes básicos da carga e parâmetros são examinados e resolvidos.

Na etapa de "Coleta de Dados" todos os dados que foram identificados como importantes para o modelo devem ser coletados. É realizada através da monitoração dos sistemas e seus componentes, idealmente com a utilização de ferramentas específicas, tais como, monitores de desempenho, sistema de contabilidade de *logs*, sistemas de gerenciamento e outros.

Os dados coletados são analisados na etapa de "Caracterização da Carga" para determinar o intervalo específico de medição a ser utilizado no estudo. Vários métodos são usados para examinar os valores e distribuições dos parâmetros. Esta etapa também pode ser usada para detectar problemas potenciais no sistema e para estudar a sua operação sob a carga corrente.

A próxima etapa é a "Criação do Modelo". Um modelo de carga de trabalho é a representação que imita o estado real da carga. Aqui é realizada a substituição da massa de dados coletados para cada componente anteriormente identificado por medidas representativas, através de refinamento e sumarização. Os valores obtidos são então atribuídos aos parâmetros selecionados, obtendo-se o modelo de carga.

Na construção de qualquer modelo, abstrações da carga são realizadas. Estas abstrações comprometem a exatidão do modelo, de modo que ele deve ser validado dentro e uma margem aceitável de erros, num processo denominado "Validação do Modelo" de carga. Esta fase é realizada executando uma carga de trabalho sintética e comparando-a com as medidas de desempenho obtidas com os resultados na execução da atual carga de trabalho. Segundo MENASCÉ & ALMEIDA (1998), se os resultados forem comparados e uma margem de erro de 10 a 30% for encontrada, o modelo de carga de trabalho pode ser considerado válido. Caso contrário, ele deve ser refinado para representar com mais exatidão a carga de trabalho atual.

Para realizar esta validação da maneira relatada na literatura é necessário realizar uma avaliação conforme descrito na figura a seguir. Entretanto verifica-se que na maioria dos casos, na prática, isto é muito difícil, e algumas vezes até impossível. Não tem como parar o ambiente real para rodar uma carga sintética e fazer a avaliação da mesma.

Como opção deve-se então utilizar o conhecimento e *feeling* dos responsáveis pelos ambientes e validar a carga sintética através de outros meios. Como exemplo, pode-se citar, uma comparação entre o percentual de utilização de recursos entre o ambiente real e o modelo.

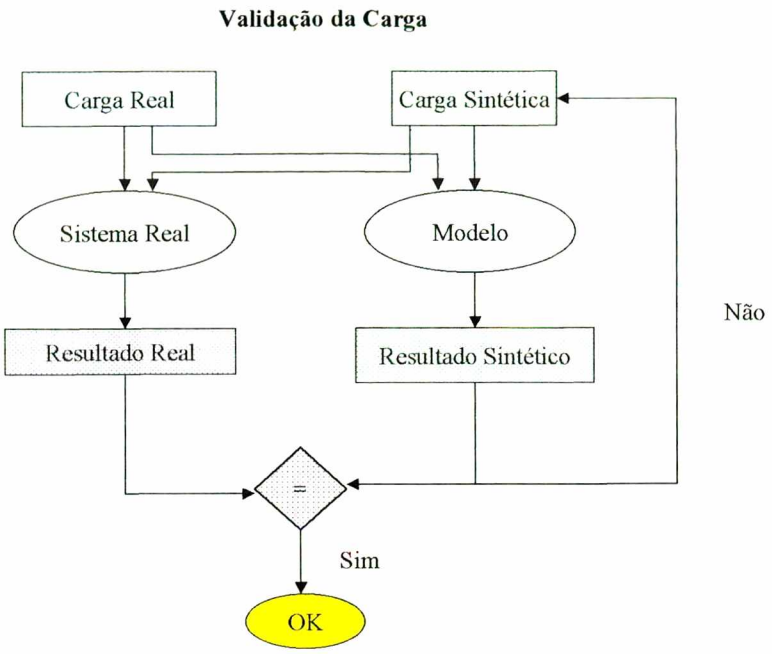


FIGURA 3 – Passos para validação da carga de trabalho

2.1.2.2.2 Nível de detalhamento

O sistema deve ser visto como um provedor de serviços a serem listados. A escolha do nível de detalhamento abrange o registro das solicitações de serviços. Existem várias possibilidades de registro que são apresentadas a seguir:

- **Requisições mais Freqüentes:** embora não possa fornecer informações suficientes sobre o sistema, é geralmente utilizado como carga inicial. São particularmente válidas, se um tipo de serviço é requisitado com muito mais freqüência do que os outros ou se é o principal consumidor dos recursos no sistema;

- **Freqüências por tipos de Requisições:** requer a lista de vários serviços com suas características e freqüência;

- **Seqüência das Requisições no Tempo:** sugere a coleta dos registros de solicitações de acordo com a seqüência cronológica em que foram ocorrendo num sistema real e o consideram como carga. Esta alternativa pode ser muito detalhada, o que a torna inconveniente para a modelagem analítica e também muito complexa para a simulação, pois requer uma reprodução exata do comportamento dos componentes para manter os relacionamentos temporais;

- **Demanda média do recurso:** utiliza a demanda média do recurso no lugar das requisições para compor a carga de trabalho.

2.1.2.2.3 Representatividade

Um modelo de carga deve ser uma representação da aplicação real. Para que haja representatividade é necessário que o teste da carga e da aplicação real sejam semelhantes nos seguintes aspectos:

- **Taxa de Chegada:** a taxa de chegada deve ser idêntica ou proporcional àquela da aplicação;

- **Demanda de Recursos:** a demanda total em cada um dos recursos essenciais deve ser idêntica ou proporcional àquela da aplicação;

- **Perfil de Utilização dos Recursos:** refere-se à seqüência e à quantidade na qual diferentes recursos são utilizados em uma aplicação.

2.1.2.2.4 Atualidade

Manter as medições atualizadas é tarefa árdua. O problema é que o comportamento do usuário real não é constante. Cada comportamento do usuário observado aplica-se a ambiente, sistema e tempos específicos. O usuário muda o seu

perfil com o tempo e qualquer mudança de desempenho faz com que o usuário mude o seu comportamento de utilização.

2.1.2.3 Desenvolvimento do modelo de desempenho

Consiste em tentar representar o sistema do mundo real, que é composto por um conjunto de recursos de *hardware* e *software*, geralmente atendidos por filas, pelos quais inúmeras entidades ou transações competem visando realizar os serviços existentes. Este modelo é usado para computar as métricas de desempenho como tempo de resposta, *throughput*, e utilização como uma função dos parâmetros da carga.

Existem alguns métodos para avaliação e modelagem de sistemas, conforme abaixo:

- **Monitoração:** somente é possível se já houver um modelo existente. Pode gerar grandes dificuldades no momento da interpretação dos dados adquiridos;

- **Modelo analítico:** indicado para situações em que as medidas são mais difíceis de serem obtidas, ou para modelos que ainda não existam, ou ainda em projetos que se baseiam nos modelos existentes. O uso de modelagem analítica permite testar modificações no modelo existente a um baixo custo operacional, validando hipóteses antes de suas implantações;

- **Modelos computacionais:** a intenção principal é a imitação do sistema real em que entidades, representando transações, fluem através de uma rede de interconexões na busca dos recursos necessários para a realização de seus processos. Controlando este fluxo, cria-se uma "história artificial" do sistema modelado. O emprego deste tipo de modelagem exige a criação de um programa. Atualmente são inúmeras as ferramentas e ambientes computacionais voltados a este tipo de modelagem.

Após o desenvolvimento do modelo, é necessário efetuar a validação. Ela consiste em assegurar que o modelo e os inúmeros pressupostos e simplificações adotados, no seu desenvolvimento, sejam razoáveis e, se corretamente implementados, tenham um comportamento e produzam resultados semelhantes àqueles observados nos sistemas reais.

2.1.2.4 Previsão de Carga de Trabalho

Esta etapa objetiva estimar o comportamento futuro da carga sobre o sistema. A idéia central é que se possa prever o desempenho do sistema, considerando as diversas aplicações diante de possíveis cenários de demanda. Isso implica conhecer a tendência de crescimento da demanda para os vários serviços disponíveis junto ao sistema de informações avaliado e o conhecimento dos planos de negócios futuros da empresa.

Para conhecer a tendência de crescimento é importante reconhecer padrões em gráficos de carga x tempo (gráfico do histórico da carga). Esse padrão tem forte influência no tipo de técnica que se quer adotar para a previsão de carga. Os gráficos podem mostrar padrões dos seguintes tipos:

- *Trend*: mostra a tendência a um crescimento da carga;
- Cíclico: mostra que o comportamento da carga se repete de intervalo em intervalo;
- Sazonal: similar ao cíclico, mas possui um intervalo de repetição muito menor;
- Estacionário: tendência a se comportar uniformemente com alguns desvios ao longo do tempo.

O conhecimento dos planos de negócios futuros da empresa é imprescindível, pois é a partir daí que se pode estimar o uso futuro do sistema com o auxílio de técnicas estatísticas de previsão. Basicamente as técnicas mais usadas são:

- Média variante: consiste em se prever um valor de carga para um período futuro igual à média de valores de observações prévias;
- Exponencial suave: similar a anterior, sendo que a diferença reside no fato de que são colocados mais pesos nos dados colhidos recentemente;
- Método de regressão: modelos de regressão são usados para estimar valores de uma variável (no caso a carga), como função de diversas outras variáveis. As variáveis previsíveis são chamadas dependentes e as que são usadas para prever o valor da carga são chamadas de variáveis independentes.

2.1.2.5 Previsão do Desempenho

Consiste em utilizar o modelo de desempenho e a previsão da carga para estimar os valores de determinadas variáveis, através de um conjunto de parâmetros relativos ao próprio sistema e à carga a ele submetida. Eles se dividem em parâmetros do sistema, de recursos e de carga.

2.1.2.6 Desenvolvimento e Previsão do Modelo de Custos

Envolve a determinação dos custos iniciais e dos custos de operação/manutenção. Os primeiros dizem respeito às parcelas envolvidas na compra e implementação do sistema. Os últimos tratam das despesas com manutenção, pessoal, treinamento, atualização do *hardware* e *software*, energia, segurança, telecomunicações e outros.

2.1.2.7 Análise do Custo x Desempenho

Depois de construídos os modelos de custos e de desempenho, várias avaliações de cenários e configurações podem ser feitas. Para cada cenário pode-se predizer qual o desempenho de cada componente básico da carga de trabalho global e quais serão os custos relacionados a este cenário.

A definição pelo melhor cenário representa a configuração ideal do sistema, que levará a produção dos planos de configuração, de investimentos e de pessoal. O plano de configuração especifica quais *upgrades* na plataforma de *hardware* e *software*, quais trocas na topologia de rede e arquitetura do sistema devem ser feitos. O plano de investimentos especifica o tempo necessário para efetuar os investimentos no *upgrade* descrito no plano anterior. O plano de pessoal especifica quais mudanças devem ocorrer na estrutura de pessoal hoje existente para absorver as trocas no sistema.

2.1.3 Dificuldades no planejamento de capacidade

Existem alguns pontos de dificuldade que surgem no processo de planejamento de capacidade:

- Necessidade de conhecer bem todo o ambiente que o processo será aplicado;
- Realização de avaliações da carga de trabalho periodicamente;
- Dificuldade em prever alterações das tecnologias;
- Entendimento do desempenho requerido pelos usuários;
- Previsão da carga de trabalho futura;
- Avaliação de uma configuração futura;
- Gerenciamento contínuo do processo.

2.2 Avaliação de desempenho

A avaliação de desempenho compreende comparar diferentes alternativas de configurações de sistemas com o objetivo de encontrar aquela que melhor satisfaz as exigências da relação custo x beneficio. Pode-se afirmar que o objetivo principal na avaliação de um sistema computacional é obter o maior desempenho com um determinado custo.

A avaliação de desempenho é também considerada uma "arte", pois avaliações bem sucedidas não podem ser realizadas mecanicamente. Cada uma delas requer um

conhecimento profundo do sistema que está sendo observado e uma escolha cuidadosa da técnica de avaliação, carga de trabalho e ferramentas que serão utilizadas.

2.2.1 Objetivos

A avaliação de desempenho envolve técnicas e metodologias que auxiliarão na resolução de problemas, como:

- Especificação de requisitos de desempenho;
- Avaliação de projetos alternativos;
- Comparação de dois ou mais sistemas;
- Determinação de valores de parâmetros ideais;
- Procura do gargalo de desempenho;
- Caracterização da carga de um sistema.

2.2.2 Metodologia de avaliação de desempenho

No momento em que se define uma metodologia de avaliação de desempenho, deve-se tomar muito cuidado para não incorrer em erros comuns, como: falta de objetivos, objetivos tendenciosos, técnica errada de avaliação e detalhamento demasiado ou inexistente. JAIN (1991) propõe uma abordagem sistemática, conforme descrito a seguir.

2.2.2.1 Definição dos objetivos e do sistema

O primeiro passo em qualquer projeto de avaliação de desempenho é a determinação dos objetivos a serem estudados e a definição do que consiste o sistema, delineando os limites do mesmo. Os objetivos podem ser a princípio difíceis de precisar, mas são essenciais para a resolução do problema, assim como a definição exata das

"fronteiras" do sistema. Estes fatos afetarão as métricas de desempenho, bem como as cargas usadas para a comparação.

2.2.2.2 Elaboração da lista de serviços e resultados esperados

Cada sistema provê uma lista de serviços e para cada um deles existe um conjunto de possíveis resultados, desejados, ou não. Por exemplo, um sistema de base de dados pode responder a uma consulta correta, incorretamente ou não responder. Uma lista dos serviços e possíveis conseqüências é importante para selecionar as métricas corretas e a carga de trabalho.

2.2.2.3 Seleção das métricas

Métricas são critérios para a comparação do desempenho. Não há uma definição padrão das métricas inseridas no contexto de avaliação de desempenho, sendo que elas dependem basicamente do comportamento dos componentes do sistema a ser estudado. De forma geral são associadas aos três tipos de resultados possíveis de uma solicitação de serviço:

- **Solicitação atendida corretamente**: neste grupo incluem-se as métricas relacionadas ao tempo usado para realizar o serviço, a taxa em que ele é realizado e os recursos utilizados enquanto é executado. Estas três medidas tempo-taxa-recurso para um desempenho bem sucedido são também chamadas de medidas de "rapidez", produtividade e utilização, respectivamente. Por exemplo, a rapidez de um *gateway* de rede é medido pelo tempo de resposta – tempo entre a chegada de um pacote e a sua correta remessa. A produtividade é medida por sua taxa (*throughput*) – o número de pacotes enviados por unidade de tempo. A utilização dá uma indicação da percentagem de tempo que os recursos do *gateway* estão ocupados. O recurso com a mais alta utilização é chamado de gargalo. Encontrar a

utilização dos vários recursos dentro do sistema é uma parte importante na avaliação de desempenho;

- **Solicitação atendida incorretamente**: neste grupo se incluem as métricas referentes à confiabilidade do sistema;

- **Solicitação não atendida**: neste grupo incluem-se as métricas ligadas à disponibilidade do sistema. É importante classificar os erros ou falhas e determinar a probabilidade de ocorrência dos mesmos.

As métricas mais comuns são:

- Tempo de resposta: é definido como o intervalo de tempo entre a requisição e a resposta fornecida pelo sistema;

- *Throughput:* é definido como a taxa (requisições por unidade de tempo) que cada requisição pode ser executada pelo sistema. Cresce à medida que a carga de trabalho aumenta até que se atinja um limite, o qual é chamado de capacidade nominal do sistema;

- Utilização de um recurso: é medida como a fração de tempo em que o recurso esteve ocupado resolvendo o serviço requisitado;

- Confiabilidade: é medida pela probabilidade de ocorrência de erros ou pelo tempo entre ocorrências de erros;

- Disponibilidade: é o tempo em que o sistema esteve ou fica disponível para atender às requisições de serviços;

- Eficiência;

- Produtividade;

- Relação custo/desempenho.

É importante, quando se está avaliando as métricas, verificar se elas são globais ou individuais. As métricas individuais refletem a utilização do sistema por um único usuário, enquanto a global reflete a utilização do sistema como um todo. A utilização, confiabilidade e disponibilidade são métricas globais, enquanto o tempo de resposta e o *throughput* são métricas que podem ser tanto individuais como globais.

Definidas as métricas, é importante considerar os itens a seguir:

- Baixa variabilidade: reduz o número de repetições para se conseguir o nível de confiança estatístico desejado;

- Não redundância: se duas métricas dão a mesma informação, é menos confuso estudar somente uma delas. Porém isto não é sempre tão óbvio;

- Conjunto completo de métricas: todos os resultados possíveis devem ser incluídos no conjunto de métricas de desempenho.

2.2.2.4 Elaboração da lista de parâmetros

O próximo passo no projeto de desempenho é fazer uma lista de todos os parâmetros que afetam o desempenho. A lista pode ser dividida em dois tipos: parâmetros de sistema (que geralmente não variam de uma instância de sistema para outra) e de carga (que são características das solicitações dos usuários e portanto bastante variáveis).

Os parâmetros de sistema são características relacionadas ao desempenho. Exemplos de parâmetros relacionados ao desempenho de sistemas cliente/servidor são

protocolos de rede, número máximo de conexões suportadas por um servidor web ou o número máximo de consultas suportado por um sistema gerenciador de base de dados.

Os parâmetros de carga são aqueles derivados da caracterização da carga de trabalho submetida ao sistema. Subdivide-se em dois segmentos:

- Parâmetros de intensidade de carga: é a medida de carga submetida ao sistema, indicada pelo número de trabalhos (requisições, comandos, transações) que disputam os recursos do sistema. Exemplo: número de buscas/dia ao servidor *proxy*, número de requisições/seg. submetidas ao servidor de arquivos;

- Parâmetros de demanda de carga: são os valores que especificam as necessidades de serviços por cada componente básico sobre cada recurso. Exemplo: tempo de CPU necessário a uma transação em um servidor de base de dados, tempo de transmissão sobre uma LAN de respostas emitidas por um servidor web.

2.2.2.5 Seleção dos fatores para desempenho

Fatores são os parâmetros que quando variados vão influenciar com mais intensidade o desempenho do sistema. Os valores que eles podem assumir são chamados de níveis. Para facilitar é melhor começar com menos fatores e poucos níveis em cada um e ir aumentando a lista conforme a necessidade. Para boa escolha dos fatores devem-se usar os parâmetros que mais influenciam no desempenho.

Na escolha dos fatores é importante considerar economia, política e limitações tecnológicas que possam existir, bem como as limitações impostas pelo responsável pelas decisões e o tempo disponível para a tomada das mesmas. Isto aumenta a chance de achar uma solução aceitável e implementável.

2.2.2.6 Seleção da técnica de avaliação

As técnicas de avaliação são a simulação, a modelagem analítica e a medição. Existem várias considerações que ajudam a decidir qual técnica usar. Elas são mostradas e ordenadas na tabela 1 da que apresenta maior para a da menor importância.

Critério	Mod. Analítica	Simulação	Medição
1. Etapa	Qualquer	Qualquer	"Protótipo Final"
2. Tempo Disponível	Pequeno	Médio	Variável
3. Ferramentas	Analistas	Linguagens Comp.	Instrumentação
4. Precisão	Baixa	Moderada	Variável
5. Equilíbrio de parâmetros	Fácil	Moderado	Difícil
6. Custo	Pequeno	Médio	Alto
7. Aceitabilidade	Baixa	Média	Alta

TABELA 1 – Critérios para seleção de uma técnica de avaliação

A principal consideração é a fase do ciclo de vida em que o sistema se encontra. Medição somente é possível se algo similar ao sistema proposto já exista. Se for projeto novo, somente pode-se escolher por modelagem analítica ou simulação.

A próxima consideração é o tempo disponível para se fazer a avaliação. Na maioria dos casos os resultados são requeridos para "ontem". Se for esse o caso, a modelagem analítica é provavelmente a única escolha. Simulações tomam bastante tempo.

O nível de precisão desejado é outra consideração importante. Geralmente a modelagem analítica requer muitas simplificações e suposições. Simulações podem incorporar mais detalhes e requerem menos suposições e freqüentemente estão mais próximos da realidade. Medições, apesar de soar como mais próximas da realidade, não podem gerar resultados precisos simplesmente porque parâmetros, tais como configurações do sistema, tipos de cargas de trabalho e tempo de medição, podem ser únicos para o experimento.

O objetivo de todo estudo de desempenho é também comparar diferentes alternativas para encontrar um valor ótimo. Modelos analíticos geralmente têm a melhor visão sobre o efeito da interação entre os parâmetros. Com a simulação é possível buscar a melhor combinação de valores dos parâmetros. Mas freqüentemente não fica clara a relação de compensação existente entre os parâmetros. Medição é a técnica menos desejável nesse sentido.

O custo destinado ao projeto é bastante importante. A medição exige instrumentos e tempos reais, e é a mais cara das três técnicas. Simulação é uma boa alternativa pela facilidade de alteração de configurações.

Segundo JAIN (1991), escolher qual das três técnicas usar depende do tempo e dos recursos disponíveis para a resolução do problema e nível de acuidade desejado. O ideal é que, independentemente da técnica escolhida, usem-se as outras para corroborar o resultado.

2.2.2.7 Seleção da carga

A carga consiste em uma lista de solicitações de serviço ao sistema e deve refletir o seu uso real. A seleção da carga deve levar em consideração o seguinte:

- Deve "exercitar" todos os serviços que importem ao estudo, ou seja, todos os serviços que afetem o desempenho devem ser solicitados;
- O nível de detalhe deve refletir a realidade, ou seja, se o sistema recebe uma grande variedade de solicitações; o uso de apenas uma como carga não é representativa;
- Parâmetros como taxa de chegada de solicitações, uso de recursos, seqüência e quantidade de uso devem também estar o mais próximo possível do uso real;
- A carga deve representar a utilização atual do sistema.

Os valores máximos e médios de carga devem ser analisados considerando-se todo o período de tempo que são amostrados. Assim pode-se ter noção da influência de períodos críticos para um sistema *on line* se considerados os horários de pico.

2.2.2.8 Planejamento dos experimentos

De posse da lista de fatores e seus níveis, deve-se decidir uma seqüência de experimentos de modo a obter o máximo de informações com o mínimo de esforço. Na prática, recomenda-se dividir os experimentos em duas fases:

- Número de fatores alto e o número de níveis mais baixos;
- Número de fatores reduzido e o número de níveis daqueles que são significativos é aumentado;

2.2.2.9 Análise e interpretação dos dados

É importante reconhecer que os resultados das avaliações e simulações são quantidades sem método e podem ser diferentes a cada experimento repetido. Na comparação de alternativas é necessário ter que considerar a variabilidade dos resultados. A simples comparação de médias pode levar a resultados insatisfatórios. A interpretação de resultados de uma análise é a chave da arte de analisar. Deve ser entendido que análises somente produzem resultados e não conclusões. Eles providenciam a base para que os analistas ou tomadores de decisão possam extrair conclusões.

2.3 Proposta de alteração da metodologia

Analisando a metodologia de planejamento de capacidade proposta por MENASCÉ & ALMEIDA (1998) observam-se alguns pontos que não estão bem especificados. Em especial o item "Previsão de Desempenho". Propõe-se a seguir uma

nova seqüência de passos, sugerindo uma interação entre as metodologias descritas anteriormente, de maneira a se complementarem. A tabela a seguir apresenta os passos de cada metodologia.

Planejamento de Capacidade	Avaliação de Desempenho
Compreensão do Ambiente	Definição dos Objetivos e do Sistema
Caracterização da Carga de Trabalho	Elaboração Lista Serviços e Resultados Esperados
Desenvolvimento Modelo de Desempenho	Seleção das Métricas
Previsão da Carga de Trabalho	Elaboração da Lista de Parâmetros
Previsão do Desempenho	Seleção dos Fatores para Desempenho
Criação do Modelo de Custos	Seleção da Técnica de Avaliação
Previsão do Custo	Seleção da Carga
Análise do Custo x Benefício	Planejamento de Experimentos
	Análise e Interpretação dos Dados

TABELA 2 – Comparativo das Metodologias

A tabela a seguir descreve a proposta de adaptação da metodologia.

Metodologia de Planejamento de Capacidade Proposta
Compreensão do Ambiente
Definição dos Objetivos e Negócios
Caracterização da Carga de Trabalho
Desenvolvimento do Modelo de Desempenho
Previsão da Carga de Trabalho
Seleção das Métricas
Seleção dos Fatores para Desempenho
Planejamento dos Experimentos
Criação e Previsão do Modelo de Custos
Análise do Custo x Desempenho

TABELA 3 – Metodologia de Planejamento de Capacidade proposta

2.3.1 Justificativa

O item "Definição dos objetivos e do negócio" não está contemplado na proposta de planejamento de capacidade. Entretanto justifica-se a sua necessidade pelo fato de que para a realização de qualquer projeto, qualquer planejamento, faz-se necessário saber exatamente onde se quer chegar. Qualquer trabalho sem objetivos está fadado a falhar, portanto é importante estabelecer inicialmente quais as respostas que o sistema deve apresentar em termos de qualidade de serviço. As métricas, cargas de trabalho e metodologias, todas dependem do objetivo. Uma vez que o problema é claro e os objetivos foram bem escritos, encontrar a solução é freqüentemente mais fácil.

Outro fator de suma importância é o conhecimento dos negócios e dos planos estratégicos da organização para o correto dimensionamento do sistema com o objetivo fim, permitindo planejar a capacidade excedente sem negligenciar a necessidade de rentabilidade.

A etapa de "Previsão de Desempenho" proposta por Menascé consiste em utilizar o modelo de desempenho e a previsão da carga para estimar os valores de determinadas variáveis no ambiente para a definição de cenários. Contudo não esclarece como isto pode ser realizado. Sugere-se então a substituição desta pelas etapas "Seleção das métricas", "Seleção dos Fatores para Desempenho" e "Planejamento de Experimentos" da metodologia de avaliação de desempenho. Isto é justificável, pois, neste ponto faz-se necessário escolher entre vários cenários possíveis. Então o ideal é fazer uma avaliação para verificar qual destes cenários apresenta o melhor desempenho em função das métricas definidas. Seguindo estas etapas, os melhores cenários podem ser escolhidos com uma base técnica realizada através de um projeto experimental.

Em seguida, juntamente com o administrador do sistema, escolhem-se os melhores cenários em função das métricas ou aqueles com maior probabilidade de ocorrerem em função do ambiente externo e da conjuntura atual. Escolhidos os cenários, realiza-se o modelo de custos somente para os mesmos. Na seqüência, os cenários que

apresentam uma melhor relação custo x benefício são os candidatos em potencial para implantação.

Um novo desenho da metodologia fica:

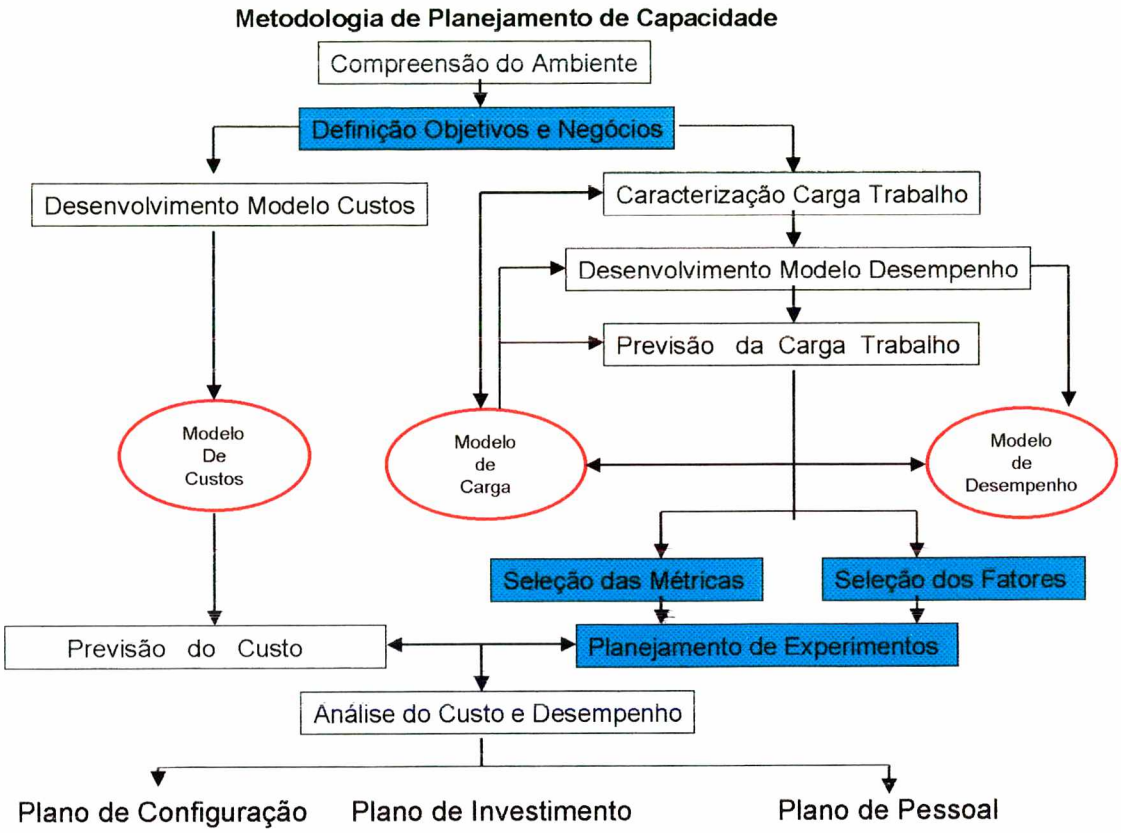


FIGURA 4 – Proposta de Metodologia de Planejamento de Capacidade

2.4 Considerações finais

O planejamento de capacidade de um sistema é definido em termos de serviços e esses são ditados pelas necessidades do negócio. As cargas decorrentes de mudanças de negócio não podem ser realocadas de uma hora para outra e nem o tempo de resposta pode ser diminuído sem a realização de uma mudança prévia no sistema. Portanto o

ideal é fazer um bom planejamento de capacidade de modo que, quando necessário, um aumento de configuração possa ser feito com os melhores requisitos técnicos e com o menor custo econômico possível.

Este capítulo mostrou a importância da realização de planejamento de capacidade e avaliação de desempenho, descrevendo-os separadamente nas suas funções e características. Ao final apresenta uma adequação da metodologia de capacidade em alguns pontos que não estão bem definidos, complementando-os com passos da metodologia de avaliação de desempenho.

No próximo capítulo é apresentada uma descrição do ambiente de provimento de serviços Internet em termos de contextualização, componentes e serviços para um embasamento da aplicação da metodologia aqui apresentada.

Capítulo 3

Ambiente de provimento de serviços Internet

Uma das grandes necessidades do ser humano sempre foi a comunicação. Com o decorrer do tempo, muitas técnicas foram inventadas para suprir tal demanda. Atualmente existe um meio de comunicação que vem superando expectativas a cada ano. É a rede mundial que interliga computadores de universidades, empresas, organizações e outras instituições, fornecendo, distribuindo e transportando qualquer tipo de informação, seja ela interativa, ou não.

Para que um usuário possa acessar a rede mundial ele necessita de uma ligação com ela. Hoje existem inúmeras instituições comerciais que provêem acesso à rede Internet. Essas instituições são chamadas Provedores de Acesso à Internet.

O ambiente escolhido para a aplicação e verificação da usabilidade e praticidade da metodologia proposta é exatamente o de Provimento de Serviços Internet. A escolha recaiu sobre este ambiente pela importância do mesmo na qualidade dos serviços oferecidos na web. Isto é justificável, pois dependendo do tráfego e capacidade instalada no provedor de serviços, problemas de desempenho podem vir a ocorrer, contribuindo para o incremento do tempo de resposta aos usuários finais.

Para uma efetiva compreensão deste ambiente este capítulo apresenta uma seção com a descrição da arquitetura Internet e suas quatro camadas. A seção seguinte contextualiza os provedores de serviços e seus componentes. Ao final, descreve os principais meios de conexão para acesso aos serviços.

3.1 Arquitetura da Internet

A arquitetura de rede Internet é um conjunto de protocolos desenvolvidos para permitir que os computadores comuniquem-se entre si em uma rede. Nesse conjunto de protocolos estão inclusos padrões que especificam os detalhes de como ocorre a comunicação entre os computadores, assim como convenções e normas para rotear o tráfego gerado por essa comunicação. O nome TCP/IP é também utilizado para descrever esta arquitetura devido a seus dois protocolos mais importantes: o TCP (*Transmission Control Protocol*) e o IP (*Internet Protocol*), porém existem outros protocolos que constituem essa família.

Esta arquitetura é constituída de quatro camadas que interagem entre si tornando-se flexível e adaptável a mudanças. As quatro camadas estão definidas abaixo.

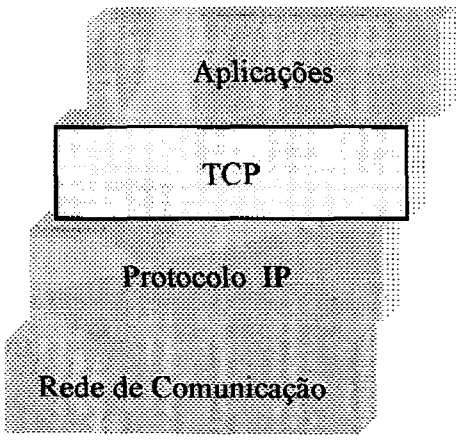


FIGURA 5 – Camadas de protocolos da arquitetura Internet TCP/IP

3.1.1 A camada interface de rede

A função principal é a interface da camada de rede da arquitetura Internet com os diversos tipos de padrões para redes (X.25, ATM, *Ethernet, Token Ring, Frame Relay,* entre outras).

3.1.2 A camada de rede

É a última camada na qual o fluxo de informação leva em conta as peculiaridades da sub-rede de comunicação, tais como sua topologia e capacidade de suas linhas físicas. A partir desta, as camadas acima empregam mecanismos de comunicação fim a fim (*host to host*), abstraindo a existência da sub-rede de comunicação. Para que a camada de transporte possa abstrair os detalhes da sub-rede de comunicação a camada de rede deve prover os seguintes serviços:

- Entrega de pacotes: recolher um pacote do *host* emissor e encaminhá-lo ao *host* receptor. Entretanto não é garantido que os pacotes sejam corretamente entregues no endereço destino;

- Roteamento: escolha da rota de comunicação por onde os pacotes oriundos da camada de transporte irão trafegar;

- Controle de congestionamento: evitar que os roteadores fiquem congestionados de pacotes devido a surtos de tráfego ou roteamento mal conduzido. Para isto é utilizado o controle por descarte de pacotes;

- Interconexão de redes: possibilitar que *host* em sub-redes heterogêneas possam se comunicar (por exemplo, um *host* numa sub-rede *Ethernet* e outro numa sub-rede *Token Ring*);

Procedimentos de reconhecimento, retransmissão, ordenação e controle de fluxo são deixados para a camada superior.

3.1.2.1 Entrega de pacotes

Pacote é a unidade de informações que trafega entre os roteadores de uma rede de longa distância. Quando um *host* necessitar transferir um pacote, este o entrega a um

roteador em sua sub-rede. O roteador escolhe uma rota para o pacote tendo como destino a sub-rede ao qual o *host* de destino está conectado.

O tráfego entre os roteadores pode se dar através de um circuito virtual preestabelecido, neste caso a camada de rede é dita orientada à conexão. Se a camada de rede não suporte circuitos virtuais, a mesma é dita sem conexão ou orientada a datagrama. Via de regra as redes limitam o campo de dados dos quadros que por ela trafegam (1500 *bytes* no caso da Ethernet). Isso traz uma implicação: datagramas superiores ao tamanho máximo do quadro são fragmentados para a transmissão e remontados no destino.

Além disso, é colocado um tempo de vida máximo para o datagrama atingir o seu endereço destino. Quando um roteador detecta que o tempo do datagrama se expirou, ele simplesmente o descarta. Um dos parâmetros para a medida de tempo é dado em número de roteadores que o pacote trafegou. A função dessa imposição é evitar que datagramas fiquem circulando indefinidamente pela sub-rede de comunicação devido a problemas de falhas em roteamento.

3.1.2.2 Roteamento

O roteamento ocorre em duas situações: no estabelecimento de uma conexão em redes orientadas à conexão e em cada pacote enviado por uma rede orientada à datagrama. Na rede Internet, a existência de roteadores obriga, na transmissão de um pacote, a se proceder a escolha de um circuito entre o emissor e o destinatário do pacote. No circuito, um ou mais roteadores irão receber o pacote, armazená-lo temporariamente e retransmiti-lo ao próximo roteador do circuito.

Os roteadores são depósitos de pacotes e somente o roteador diretamente conectado à sub-rede do *host* destinatário está habilitado a fazer a entrega. Em redes de computadores não se dispõe de informações antecipadas quanto ao fluxo de pacotes. A forma mais simples de rotear pacotes numa sub-rede é dotar cada roteador de uma tabela de roteamento contendo a sub-rede de destino e o próximo roteador envolvido.

Ainda existe a possibilidade do armazenamento de mais de um destino na eventualidade de uma via de comunicação ou um roteador do circuito sair de serviço. Num roteamento dinâmico, a tabela de roteamento é atualizada periodicamente em função das condições de tráfego e da mudança de topologia da sub-rede de comunicação.

3.1.2.3 O protocolo IP

O protocolo IP (*Internet Protocol*) é a base da rede Internet. A interconexão de redes na arquitetura TCP/IP supõe que todas as sub-redes sejam capazes de manipular datagramas (pacotes) padronizados. O protocolo IP fornece exatamente um padrão para a construção e manipulação de datagramas que irão circular pelas sub-redes de comunicação.

3.1.2.4 Endereçamento IP

Cada interface de rede associada a um roteador ou a uma máquina que esteja conectada à Internet pode possuir um número IP. Esse número IP guarda informações de sua rede (*network*) e do seu número de máquina (*host*). Essa combinação é dita única, ou seja, não podem existir duas máquinas na Internet com o mesmo número IP.

3.1.3 A Camada de transporte

Localizada acima da camada de Rede, esta camada permite que entidades na fonte e no destinatário entrem em conversação.A arquitetura TCP/IP tem dois principais protocolos na camada de transporte: um orientado à conexão (*Transmission Control Protocol* – TCP) e outro não orientado à conexão (*User Datagram Protocol* – UDP). O protocolo UDP é basicamente igual ao protocolo IP acrescentado de um pequeno *header;* já o TCP é bem mais elaborado.

3.1.3.1 O funcionamento dos protocolos TCP/UDP

Um serviço TCP ou UDP é obtido, colocando-se o receptor e o emissor ligados a pontos chamados de *sockets*. Um *socket* determina unicamente um ponto de ligação em uma máquina, que tanto pode ser de origem como destino. Cada *socket* tem um número associado a ele, chamado de *socket number*, que é constituído do número IP da máquina, acrescentado de um número de 2 *bytes* interno ao *host*, denominado porta (*port*). Para obter um serviço TCP ou UDP uma conexão deve ser explicitamente estabelecida entre o *socket* emissor e o *socket* do *host* receptor.

Fonte: TOPKE, Claus Rugani, Provedor Internet - Arquitetura e Protocolos. São Paulo: MAKRON Books. 1999. 1ª edição.

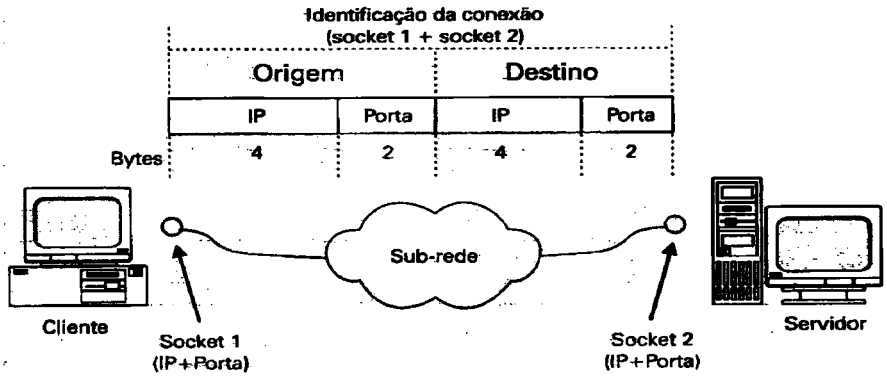


FIGURA 6 – Uma conexão utilizando *sockets*

Para facilitar as conexões a servidores foram reservadas as portas de números até 1024. Assim um programa cliente pode utilizar um serviço do servidor, bastando para isso se conectar a tal porta. Como, por exemplo, um cliente necessitando se conectar a um servidor WWW se conecta à porta 80 deste. Um processo que precisa se conectar a um servidor para transferir arquivos através de um protocolo FTP necessita faze-lo na porta 21 do servidor.

O processo de uma conexão começa na criação de um *socket* no cliente, que geralmente utiliza um número de porta não reservado, ou seja, maior do que 1024. Após a criação do *socket* no cliente, este tenta se conectar ao servidor, que necessariamente

deve possuir um *socket* já criado. Este *socket* no servidor é criado pelo programa servidor, o qual espera por conexões de clientes. O funcionamento de um *socket* cliente e de um *socket* servidor são diferentes: no servidor o *socket* é "escutado" pelo programa servidor, que pode receber pedidos de conexão de mais de um *socket* cliente; e no cliente é apenas criado para o estabelecimento da conexão durante a necessidade e depois ele é destruído.

3.1.3.2 O protocolo TCP

É um protocolo confiável orientado à conexão, o que permite que o envio de uma seqüência de dados originados em uma máquina, seja enviado sem erros, para qualquer outra máquina na Internet. Esta camada, se necessário, fragmenta uma mensagem grande em pedaços de mensagens pequenas e os envia para a camada Internet. No destinatário, o processo TCP de recepção remonta as pequenas mensagens no tamanho original. O TCP ainda fornece controle de fluxo para garantir que um "transmissor" rápido não atropele um receptor "lento", o qual não pode tratar tão rápido a recepção das mensagens.

3.1.3.3 O protocolo UDP

É um protocolo não confiável e não orientado à conexão. É utilizado por aplicativos que não desejam o controle de fluxos e a ordem de entrega que o TCP objetiva, podendo usar os seus próprios. É mais utilizado por programas cuja necessidade de tempo de resposta é maior que a necessidade de garantia de chegada e de ordem, como em uma conversação ou em vídeo.

3.1.4 A camada de aplicação

Esta camada define um conjunto de serviços manipulados diretamente pelo usuário; tais serviços utilizam protocolos definidos nesta camada. Os serviços podem, ser ou não, transparentes ao usuário final.

Na arquitetura TCP/IP, os serviços da camada de aplicação utilizam a filosofia cliente-servidor. Os aplicativos são clientes dos serviços definidos pelas implementações já existentes no TCP/IP. Clientes e servidores se comunicam através de protocolos de aplicação que regem a interação cliente-servidor. Grande parte dos protocolos de aplicação se utilizam dos protocolos de transporte TCP e UDP para transmissão de suas mensagens. Alguns dos protocolos de alto nível desta camada estão descritos a seguir.

3.1.4.1 Terminal remoto

É um serviço através do qual um terminal conectado a um computador se apresenta ao usuário como se estivesse conectado a outro computador remoto. O protocolo Telnet é um dos padrões de comunicação entre cliente e servidor de terminal remoto. O principal problema enfrentado por esse tipo de protocolo é a falta de segurança. Em substituição ao mesmo, vem sendo utilizado o protocolo SSH (*Secure Shell*), para conexões de *login* remoto seguras usando meios de encriptação e autenticação. Sua porta é a 23 e há várias RFC's sobre ele, como, por exemplo, 1205, 1198, 1091 entre outros.

3.1.4.2 *File Transfer Protocol* (FTP)

Permite que um usuário em um computador transfira, renomeie ou remova arquivos remotos, ou crie, remova e modifique diretórios remotos. Definido na RFC 959, duas conexões de transporte são estabelecidas: uma utilizada para interação cliente-servidor (denominada conexão de controle) e outra exclusiva para transferência de arquivos (denominada conexão de dados). A conexão de controle permanece aberta enquanto durar a sessão FTP. Para estas comunicações se faz uso das portas 20 e 21.

3.1.4.3 *Simple Mail Transfer Protocol* (SMTP)

O correio eletrônico permite a um usuário, numa determinada máquina, enviar mensagens a outro usuário em qualquer lugar da rede Internet. Para isso se utiliza do

protocolo SMTP. Cada usuário dispõe de uma área de armazenamento temporário de mensagens, denominada *main spool* para a recepção. E uma área em conjunto utilizada para enfileirar as mensagens a serem enviadas e é denominada *mail queue*. Quando o usuário envia uma mensagem através de um aplicativo, a mensagem é depositada no *mail queue* e permanece aguardando o envio. Mensagens não são obrigatoriamente enviadas assim que são submetidas. A leitura das mensagens é feita de uma forma análoga; o aplicativo utilizado verifica no *spool* se existem mensagens novas. A transferência da mensagem é executada por um processo em *background*, permitindo que o usuário remetente, após entregar a mensagem ao sistema de correio eletrônico, possa executar outras aplicações. Este aplicativo pode rodar na própria máquina ou em uma máquina remota. Para a utilização remota existem protocolos como, por exemplo, o POP.

Fonte: TOPKE Claus Rugani, Provedor Internet - Arquitetura e Protocolos. São Paulo: MAKRON Books. 1999. 1ª edição.

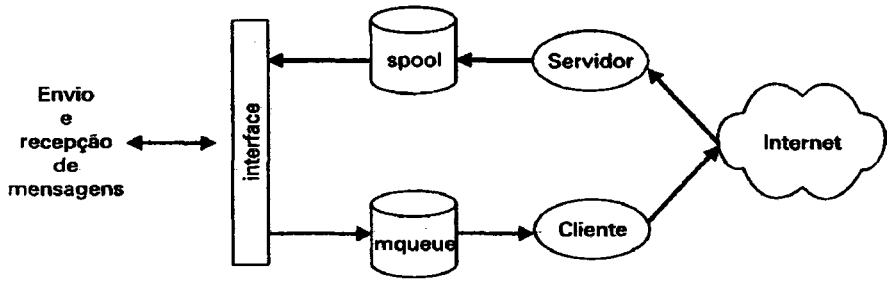


FIGURA 7 – Serviço de envio e recepção de mensagens

Na transmissão das mensagens o *host* emissor conecta-se na porta 25 (servidor SMTP) do *host* receptor. Estabelecida a conexão, cliente e servidor interagem via SMTP. Caso a conexão não possa ser estabelecida, o cliente mantém a mensagem na *mail queue* para envio futuro. Os passos na transferência de uma mensagem de correio eletrônico através do protocolo SMTP são:

- Cliente e servidor trocam *handshake;*
- O cliente identifica o emissor e o destinatário da mensagem;
- Caso o destinatário não seja reconhecido pelo servidor, a sessão se encerra com uma mensagem de informação;

- Reconhecido o destinatário, o cliente solicita que o servidor receba a mensagem;
- Por último, a sessão se encerra com o fechamento da conexão TCP de ambos os lados.

Caso o *host* de destino não possua servidor SMTP funcionando, o usuário cliente receberá mensagens informativas.

3.1.4.4 *Post Office Protocol* (POP)

Em certos nós da rede é praticamente impossível manter diretamente conectado um sistema de recepção de mensagens. Um computador pessoal pode não ter recursos suficientes para possuir seu próprio servidor de SMTP. O principal recurso necessário para a viabilização de um servidor de SMTP na máquina é a conectividade. Esse fato se deve ao próprio sistema de *mail* da Internet. Ao consultar o registro de um domínio, a máquina cliente sabe o endereço de envio para uma mensagem, então ela tenta conectar na porta 25 do *host* servidor. Se esta não estiver apta a receber tal conexão, a máquina cliente envia uma mensagem para o usuário que gerou a mensagem, dizendo que a máquina de destino não pode ser acessada. Logo, para máquinas de acesso remoto, fica impossível a implementação de um servidor SMTP.

Para a solução desse problema foi desenvolvido o protocolo POP, o qual funciona para gerenciamento remoto de *mailbox* no servidor em que se encontram as mensagens recebidas pelo SMTP. O servidor de SMTP e de POP não necessitam estar na mesma máquina. O cliente POP pode ser uma máquina qualquer; não precisa obrigatoriamente estar conectada à rede de forma permanente. A porta padrão do servidor POP é a 110, e o protocolo de transporte utilizado é o TCP.

3.1.4.5 *Hipertext Transfer Protocol* (HTTP)

É o protocolo utilizado para transferência de informações no *Word Wide Web* - WWW. Os dados transferidos pelo HTTP podem ser texto não estruturado, hipertextos,

imagens ou qualquer outro tipo de dados. O protocolo HTTP consiste em dois conjuntos de transferência de informações: o conjunto de perguntas feitas pelos clientes (*browsers*) e o conjunto de respostas feitas pelo servidor. Na versão original do protocolo HTTP 1.0, uma nova conexão é estabelecida por requisição. Na versão HTTP 1.1, também chamada de conexão persistente, uma conexão HTTP pode ser usada para transportar múltiplas requisições HTTP, eliminando o custo de abrir e fechar várias vezes.

3.1.4.6 *Domain Name System* (DNS)

Define a sintaxe de nomes usados na Internet, regras para delegação de autoridade da definição de nomes, um banco de dados distribuído que associa nomes a atributos e um algoritmo distribuído para mapear nomes em endereços. O DNS permite que qualquer endereço IP seja substituído por um nome alfanumérico, sendo que estes nomes são um pouco mais fáceis para serem lembrados e para serem aplicados em uma estrutura organizacional. O grande problema no mapeamento de nomes reside na grandeza da rede. Uma solução é a divisão hierárquica da tabela e o compartilhamento destas informações.O DNS é especificado nas RFCs 882, 883 e 973.

Na figura abaixo pode-se ver o posicionamento entre TCP, UDP e IP:

Fonte: TOPKE Claus Rugani, Provedor Internet - Arquitetura e Protocolos. São Paulo: MAKRON Books. 1999. 1ª edição.

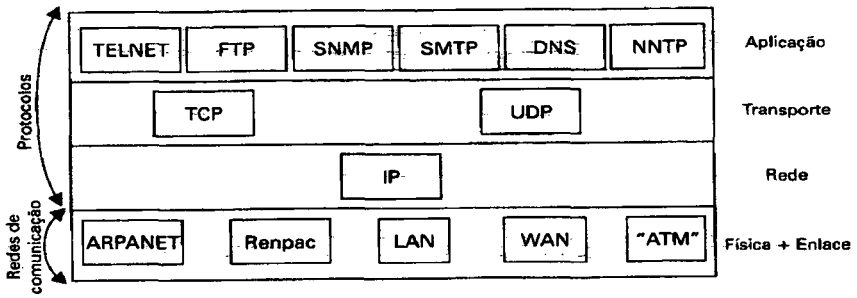


FIGURA 8 – Protocolos e redes de comunicação utilizados no modelo TCP/IP

3.2 Provedores de serviços Internet

Um Provedor de Serviços Internet - ISP é uma instituição que presta serviços Internet e se liga a um ponto de presença para obter conectividade IP e repassá-la a outras instituições, em caráter comercial ou não.

Serviços Internet são o conjunto de aplicações da camada de Aplicação (Telnet, FTP, SMTP, DNS) e serviços de informações (WWW, *Gopher, News*) que permitem a comunicação e troca de informações através da Internet.

A função de um provedor Internet pode ser caracterizada por diversos fatores, e o principal é possuir a conexão *full-time* à rede mundial. Isso pode ser conseguido através de um *backbone*. As espinhas dorsais ou *backbones* são estruturas de redes capazes de manipular grandes volumes de informações, constituídas basicamente por roteadores de tráfego interligados por circuitos de alta velocidade. As conexões dos ISP até os *backbones* são feitas através de circuitos de comunicação ponto-a-ponto, conhecidos como *links* TOPKE (1999). Os provedores Internet podem ser divididos em níveis, conforme a fig. 9.

Fonte: TOPKE Claus Rugani, Provedor Internet - Arquitetura e Protocolos. São Paulo: MAKRON Books. 1999. 1ª edição.

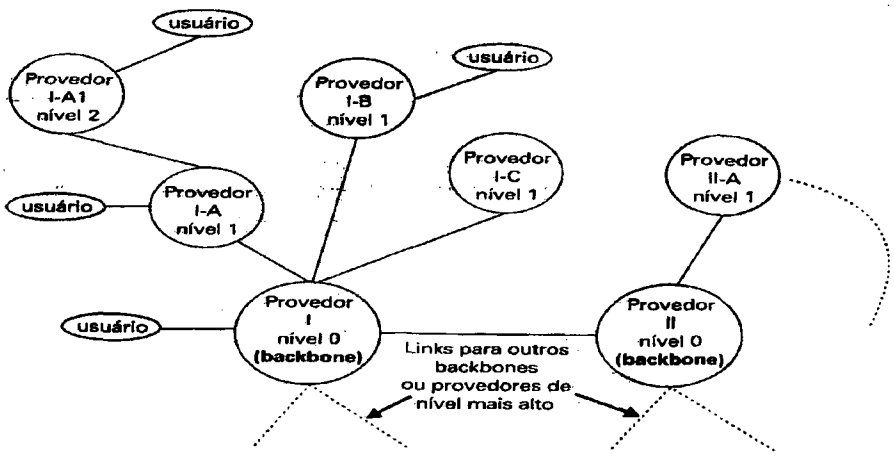


FIGURA 9 – Níveis de provedores

O nível 0 corresponde ao provedor de grande porte, sua principal característica é a de possuir ligações para outros provedores de nível 0, podendo, ou não, ter ligações com provedores de níveis mais altos. O nível 1 é o provedor que está ligado diretamente ao *backbone* (nível 0), podendo, ou não, ter ligações com provedores de níveis mais altos. O nível 2 é o provedor ligado indiretamente ao backbone através de um provedor de nível 1. Este provedor apresenta pequenas dimensões e geralmente procura este método para baratear custos. Provedores de níveis maiores podem existir, porém sua performance tende a se degradar devido ao número de *hops* em que os pacotes IP irão passar. O número de *hops* pode ser entendido como o número de roteadores que o pacote vai passar até acessar o destino. A figura abaixo mostra o *layout* de um provedor simples, em que os clientes utilizam o acesso via conexão discada para conectar-se ao provedor.

Fonte: TOPKE Claus Rugani, Provedor Internet - Arquitetura e Protocolos. São Paulo: MAKRON Books. 1999. 1ª edição.

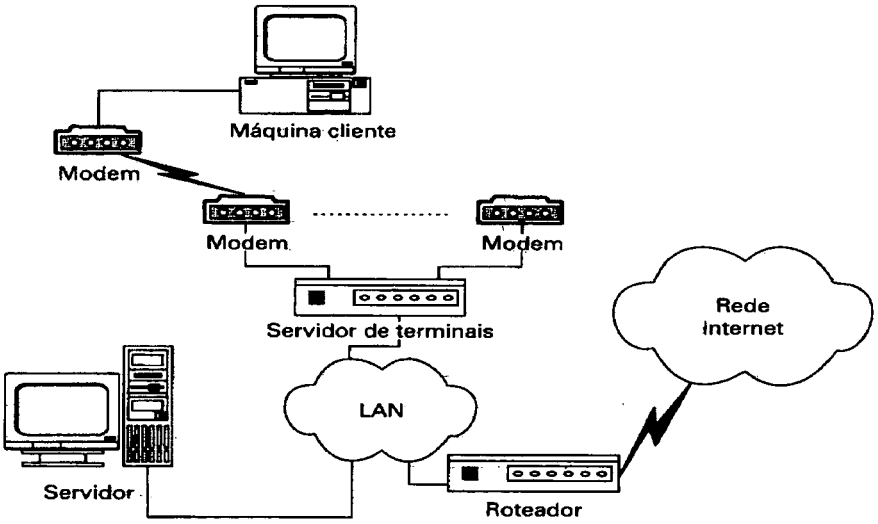


FIGURA 10 – *Layout* básico de um provedor

3.3 Componentes do ambiente de provimento de serviços

A topologia da rede no ambiente de provimento de serviços Internet envolve uma série de componentes que desempenham um importante papel na qualidade de serviço entregue ao usuário final. Os mesmos são definidos a seguir.

3.3.1 Servidor Web

É uma combinação da plataforma de *hardware*, sistema operacional, *software* de rede e servidor HTTP. Com relação ao *software* do servidor web, também conhecido como servidor HTTP, são programas que controlam o fluxo das entradas e saídas de um computador conectado a uma intranet ou a Internet. Basicamente, ele escuta o meio para verificar requisições HTTP vindo de clientes na rede. Para tanto, fica esperando conexões permanentemente em uma porta designada, geralmente a porta 80. O programa estabelece a conexão solicitada entre o servidor e o cliente, envia os arquivos requisitados e retorna para ficar em estado de escuta novamente. Do ponto de vista do *hardware*, o desempenho de servidor está em função de fatores como velocidade e número de processadores, capacidade de memória, velocidade e capacidade do sistema de disco.

3.3.2 Link

Circuitos de comunicação ponto-a-ponto com uma determinada capacidade de transmissão de dados.

3.3.3 Hub

Centro inteligente de fiação em que todos os dispositivos, impressoras, servidores, *scanners*, PC são conectados em um segmento de uma rede local. Hubs permitem que LANs sejam conectadas por pares de fios trançados em vez de cabos coaxiais. Fornecem mecanismos para localizar defeitos em máquinas e realocar dispositivos. A velocidade é geralmente de 10 Mpbs.

3.3.4 Roteador

Conectam múltiplas LANs, selecionando o melhor caminho disponível para enviar dados entre as mesmas DODD (1998). São conversores de meio, bastante utilizados em inter-redes, responsáveis por estabelecer a comunicação entre a Internet e

a rede local de computadores. Recebem os pacotes da rede, descobrem o roteamento necessário e os enviam ao destino, segundo o protocolo da rede local em que se encontra SOARES *et al* (1995). Os roteadores possuem pelo menos duas interfaces: uma compatível com a rede local e a outra ao meio em que ela se comunicará com a Internet. Geralmente essa segunda interface é do tipo síncrona, utilizando protocolos do tipo X25, *Frame Relay,* PPP, HDLC.

É importante observar que os roteadores não traduzem os protocolos. Eles simplesmente permitem que diferentes protocolos de LANs sejam transportados. O roteador permite DODD (1998):

- Controle do fluxo: se o caminho que os dados devem tomar está congestionado, o roteador pode segura-los até que o caminho entre os roteadores esteja disponível;

- Otimização do caminho: o roteador que está enviando o dado seleciona o melhor caminho disponível. As tabelas de roteamento que contêm as informações sobre os caminhos são checadas;

- Sequenciação: roteadores enviam dados em pacotes ou envelopes, que podem chegar fora de ordem no roteador final, e este os coloca na ordem correta;

- Confirmação de recebimento: o roteador final envia uma mensagem ao roteador que enviou a mensagem, informando-o de que os dados foram recebidos corretamente.

3.3.5 Gateway

Suporta arquiteturas diferentes de rede com mapeamento de endereços de uma rede para outra e também a transformação dos dados para deixar compatível entre os sistemas.

3.3.6 Firewall

É uma barreira colocada entre a rede e o mundo exterior para prevenir invasões indesejáveis e potencialmente perigosas para a rede. O objetivo é proteger um grupo de máquinas, escondendo-as de acessos não autorizados, fazendo com que todos os pacotes sejam forçados a entrar e sair apenas por um caminho, no qual podem ser inspecionados.

Fonte: SOARES, Luiz F. G.; LEMOS, Guido; COLCHER, Sérgio. Redes de Computadores. Rio de Janeiro: Editora Campus. 1995. 2ª edição.

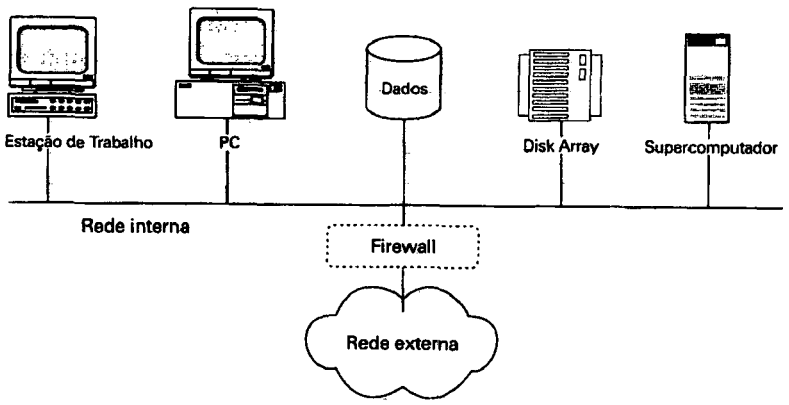


FIGURA 11 – Localização do *firewall*

Segundo SOARES *et al* (1995) o objetivo do *firewall* é garantir a integridade dos recursos ligados à rede. A centralização demanda uma administração muito cuidadosa por parte dos administradores dos sistemas, das máquinas que implementam o *firewall*. Enquanto as máquinas de uso geral são configuradas para otimizar o desempenho e a facilidade de utilização, no *firewall* tudo passa para o segundo plano, cedendo lugar ao seu objetivo principal no sistema: a segurança.

3.3.7 *Proxy* e *caches*

Proxy servers e *caches* são técnicas usadas para aumentar o desempenho, escalabilidade e segurança na web.

Cache reduz o tempo de acesso trazendo os dados o mais próximo possível dos usuários, mantendo uma cópia de objetos na própria máquina do cliente. Numa requisição HTTP, por exemplo, as figuras ficam armazenadas localmente para quando forem novamente requisitadas não necessitarem transitar pela rede. Assim *cache* melhora a velocidade de acesso e diminui o tráfego na rede, pois objetos passam a ser requisitados de lugares mais próximos. *Cache* também reduz a carga nos servidores e aumenta a disponibilidade na web.

Proxy server é um tipo especial de servidor web que trabalha como um intermediário entre a rede e a Internet, sendo capaz de agir como um cliente e como um servidor. As suas funções básicas são otimizar o desempenho da rede sob três aspectos fundamentais:

- Aumentar o número de requisições atendidas pelo servidor (o provedor consegue suportar muito mais usuários no seu *link*);

- Diminuir o volume de dados que trafegam na rede, otimizando a utilização da largura de banda (um grande número de documentos requisitados são retornados ao cliente diretamente do *proxy – byte hit ratio);*

- Reduzir o tempo de resposta para o usuário final (o *proxy* está localizado bem mais próximo dos clientes, despendendo menos tempo para enviar um objeto diretamente do *proxy* para o usuário – *cache hit ratio*).

Além de reduzir o tráfego no *link* principal do provedor, o *proxy* reduz também o tráfego geral na Internet, fazendo com que a velocidade de acesso em geral aumente. A fig. 12 mostra o mecanismo de funcionamento de um *proxy* no ambiente web:

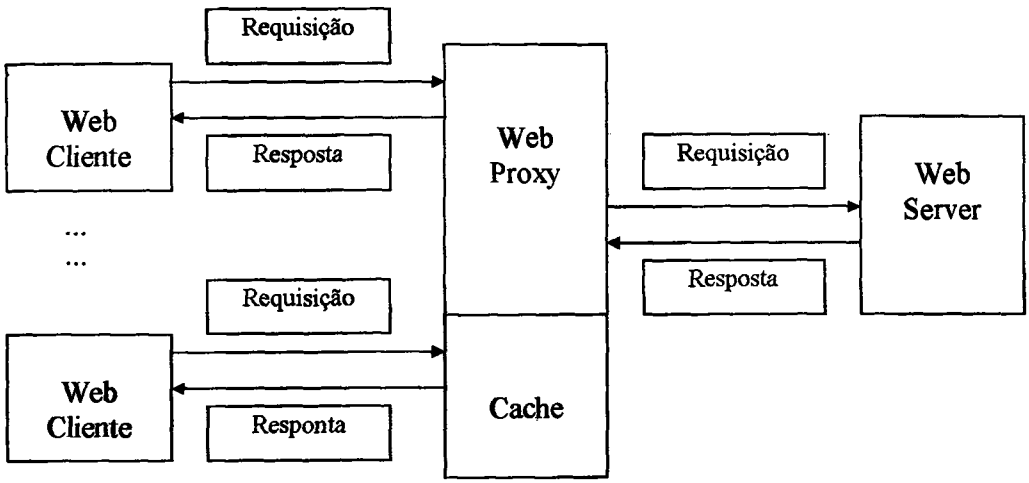


FIGURA 12 - Esquema de um *proxy server*

O funcionamento de um *proxy* pode ser explicado assim: as requisições dos usuários são direcionadas para o servidor *proxy*, que mantém cópias dos documentos mais requisitados. Quando outro usuário as solicita, ele apenas verifica se não houve modificação na página original e, caso isso se confirme, envia a página para o usuário, aumentando a eficiência da rede. Caso o documento não seja encontrado, o servidor *proxy* age como um cliente, faz uma conexão com o servidor web original, requisita o documento, armazena-o na *cache* e retorna para o cliente que o solicitou originalmente. O acesso aos documentos que são encontrados no servidor *proxy* é geralmente muito mais rápido que o acesso ao servidor original. Entretanto um documento não encontrado no servidor *proxy* gera um *delay* no tempo de resposta devido ao *overhead* da tentativa de busca no servidor.

MENASCÉ & ALMEIDA (1998) comprovaram que em média a utilização do servidor *proxy* reduz o tempo de resposta em torno de 47%. Em COCKCROFT (1997) afirma-se que a taxa de acerto varia entre 30 e 50%, fazendo com que o *link* principal possa ter sua utilização diminuída em até 50%, ou seja, a sua velocidade pode virtualmente dobrar.

As figuras abaixo ilustram os ambientes sem e com a utilização de um *proxy*:

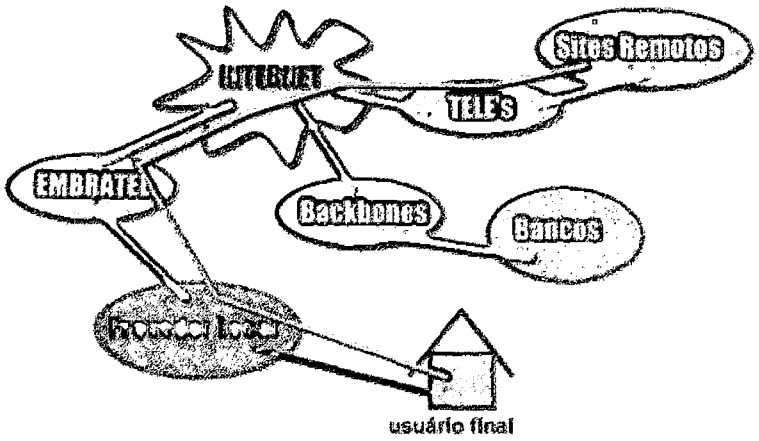


FIGURA 13 – Ambiente sem utilização de *proxy*

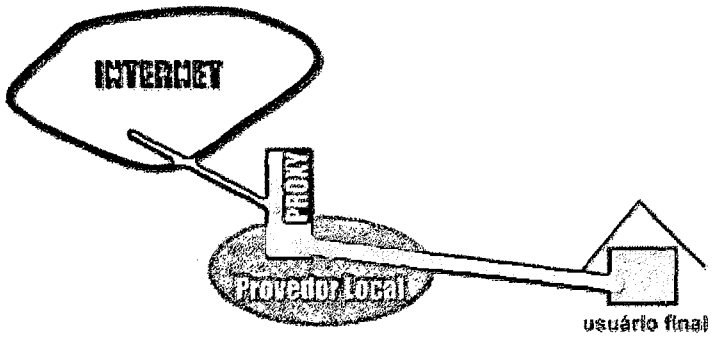


FIGURA 14 - Ambiente com utilização do *proxy*

Para usufruir das vantagens do *proxy* é necessário que o mesmo seja bem projetado. Isto envolve basicamente a determinação da melhor política de substituição de arquivos, que é utilizada quando ocorre espaço insuficiente na *cache* e é necessário escolher quais documentos devem ser removidos para liberar espaço para os novos arquivos. O funcionamento da política é assim: o gerenciador da *cache* mantém uma lista dos documentos que estão armazenados. Quando uma requisição chega, o gerenciador pesquisa na lista para localizar o documento solicitado. Se ele não está presente, o gerenciador busca o documento no servidor web original. Quando o documento chega e a *cache* está cheia, uma política de substituição de arquivos é aplicada.

As principais políticas de substituição de arquivos são:

- *Least-Frequency-used* (LFU): exclui o documento que foi acessado com menos freqüência. Isto é realizado com a manutenção de um contador de referência para cada arquivo na *cache*. Cada vez que o mesmo arquivo é acessado por um usuário o contador do arquivo requisitado é incrementado em uma unidade. Quando um documento precisa ser excluído, o arquivo com o menor contador é escolhido. Segundo ARLIT *et al* (1999) e ARLIT & FRIEDRICH (1998), 80% dos arquivos são acessados somente uma vez e apenas um pequeno número de arquivos é acessado com mais freqüência. Já YU (1998) comprovou que 34% dos arquivos no *proxy* são acessados somente uma vez e entre todos os arquivos e todas as requisições, 1% dos arquivos acessados no p*roxy* respondem por 28% de todas de requisições. MAHARTI & WILLIANSON (1999), por sua vez, afirmam que entre todos os documentos acessados aproximadamente 70% o são somente uma vez;

- LFU*: similar ao LFU com a diferença de que somente os arquivos com o contador de referência com a quantidade 1 podem ser removidos da *cache*. Entretanto há um problema nas políticas LFU e LFU* que deve ser considerado. Alguns documentos que num tempo passado foram muito acessados tiveram o seu contador de referência incrementado e dificilmente serão excluídos segundo estas duas políticas;

- LFU aging: reduz o contador dos arquivos muito requisitados para uma média entre todos os arquivos que estão armazenados;

- LFU* aging: consiste em acrescentar nas políticas anteriores dois parâmetros: *Mrefs* e *Amax*. *Mrefs* é um contador de referência máximo que um arquivo pode atingir. Se o contador de um documento na *cache* alcança este valor, o contador não será mais incrementado. Este parâmetro evita que o contador de referência de um arquivo fique muito grande. O segundo parâmetro *Amax* é o número médio máximo de referência de todos os arquivos na *cache*. Uma vez que este valor é atingido, o contador de referência de cada arquivo é reduzido para dois. Assim, se um arquivo não é mais tão requisitado como antes, eventualmente o contador de referência irá ser reduzido e ficará entre os arquivos candidatos a serem substituídos;

- *Least-recently-used* (LRU): exclui o documento que foi requisitado menos recentemente;

- *Least-Frequency Removed* (LFR): política similar ao LFU com a exceção de que os documentos que chegam na *cache* recebem um contador n+1, onde n é o número de vezes que o documento já foi removido;

- *GreedyDual-Size (*GD-Size): cada arquivo na *cache* é associado a um valor de custo H. Quando um arquivo é inicialmente armazenado o valor de custo H é setado pelo custo de trazer o arquivo até a *cache*. Quando falta espaço na *cache*, o arquivo com o menor custo H é excluído. Se o arquivo é encontrado, o valor H do arquivo requisitado é setado para o custo de trazê-lo para a *cache*;

- *GreedyDual-Size-Frequency (*GD-Size-Frequency): similar ao anterior exceto que esta política agrega o fato que a cada vez que o arquivo é referenciado, o valor H é setado para n vezes o custo de trazê-lo para a *cache* (n é o contador de referência do documento). Assim esta política integra localização temporal, custo e freqüência.

Como citado anteriormente, um *proxy server* pode potencializar o aumento do desempenho na rede sobre três aspectos: número de requisições servidas, volume de dados transferidos e tempo de resposta ao usuário final.Vários pesquisadores têm trabalhado para determinar o melhor algoritmo de *cache* em função destes aspectos. ARLIT & WILLIAMSON (1997b) propuseram várias variações do LFU e mostraram que o LFU*-Aging tem um desempenho melhor que as outras LFUs, LRU e SIZE em termos de taxa de acerto na cache e redução de volume de dados. Enquanto WILLIAMS *et al* (1996) mostraram que o algoritmo GD-Size tem desempenho melhor que LFU, LRU e suas variações em relação às duas medidas acima. ZHAO (1999) também realizou comparações entre as várias políticas acima descritas, e os resultados mostram que LFU*-Aging é o melhor algoritmo para reduzir o volume dos dados transferidos na rede, enquanto o GD-Size-Frequency apresenta o melhor desempenho em termos de maximização da taxa de acertos do *proxy* e minimização do tempo de resposta ao usuário. BUSARI (1999) realizou uma comparação entre as várias políticas e coloca que

a LFR tem melhor desempenho em *caches* que tem capacidade de armazenamento maior.

Outros pesquisadores vêm apresentando variações dos métodos acima. MURTA *et al* (1998) sugerem um gerenciamento de *cache* dividindo o espaço em classes, baseadas nos tamanhos dos objetos. Os resultados mostram que este esquema apresenta bom resultados em relação às métricas de taxa de acerto e redução de volume de dados.

Alguns trabalhos propuseram políticas que limitam os maiores e menores arquivos permitidos na *cache* ARLIT & WILLIAMSON (1997b), MARKATOS (1996). Outros pesquisadores sugeriram políticas que dividem a *cache* em várias partições, sendo que cada uma armazena um tipo de documento (como texto, imagem ou vídeo) ARLIT & WILLIAMSON (1997b) e WILLIAMS *et al* (1996). MOKHTAR & MOUFTAH (1999) sugerem uma técnica para diminuir ainda mais o tempo de resposta aos clientes através da cooperação entre *caches*.

Não foi encontrada na literatura uma política de substituição de arquivos que otimize e atenda todos os três aspectos, conforme pode ser visualizado na tabela a seguir. Neste caso, cabe a cada administrador avaliar o seu ambiente e determinar a que melhor atenda a sua finalidade específica.

Política	Aumenta Nr Req	Diminui Volume Dados(byte hit)	Reduz Tempo Resposta(cache hit)
LFU			
LFU *			
LFU aging			
LFU * aging		ARLIT & WILLIAMSON (1997b) ZHAO (1999)	ARLIT & WILLIAMSON (1997b)
LRU			
LFR			
GD-Size		WILLIAMS *et al* (1996)	WILLIAMS *et al* (1996)
GD-Size-Frequency	ZHAO (1999)		ZHAO (1999)
Dividir cache em classes		MURTA *et al* (1998)	MURTA *et al* (1998)

TABELA 4 – Comparação entre políticas de substituição de arquivos

Este trabalho não se propõe a avaliar e modelar as diferentes políticas de substituição de arquivos, entretanto haja vista a sua importância no ambiente de provimentos de serviços Internet é sabido que uma mudança nestes parâmetros pode

influenciar o desempenho do sistema como um todo. Para tanto, na etapa de experimentação é considerada uma simplificação destas políticas.

3.4 Meios de conexão

Para um usuário de computador conectar-se à rede mundial é preciso que exista um meio de comunicação entre ele e seu provedor. O tipo de acesso e serviços Internet oferecidos por um ISP podem variar bastante, dependendo de seus objetivos. Os tipos de acessos disponíveis estão descritos a seguir.

3.4.1 Conexão por discagem

Na maioria dos casos é utilizado por pequenos clientes que programam seus modems para discar para um número de telefone disponibilizado pelo provedor de serviços, que permite o acesso aos serviços Internet, através de conexões com velocidades de até 56 Kbps.

Fonte: TOPKE Claus Rugani, Provedor Internet - Arquitetura e Protocolos. São Paulo: MAKRON Books. São Paulo. 1999. 1ª edição.

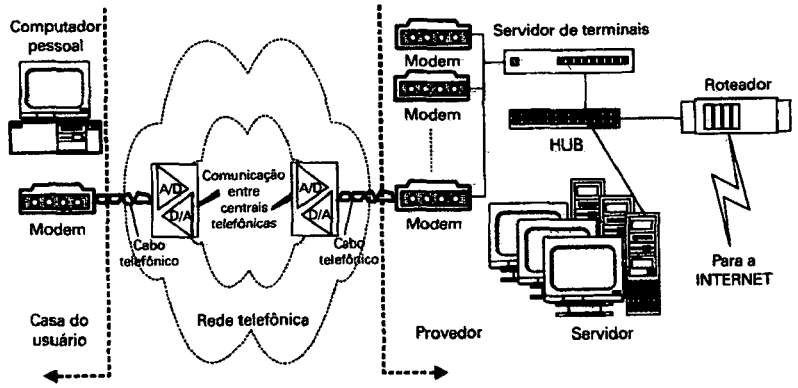


FIGURA 15 - Conexão por discagem

Os modems convertem o sinal digital recebido do computador em sinais analógicos para serem transmitidos sobre as linhas telefônicas analógicas. O processo de conversão dos sinais analógicos para digitais é chamado modulação. No receptor, o

modem demula o sinal ou converte-o de analógico para digital e o transmite para o computador. A principal limitação dos modems é o fato de eles trabalharem com linhas analógicas. A velocidade cai com o ruído nas linhas e o sinal vai decaindo em longas distâncias. Estas conexões são comuns para usuários domésticos e pequenas empresas que não necessitam de uma conexão à rede permanentemente.

3.4.2 Conexão dedicada – *Integrated Services Digital Network* - ISDN

Este serviço permite a clientes do ISP uma conexão permanente à rede e serviços Internet. A velocidade desta linha depende da negociação entre o ISP e o cliente e do suporte existente no ISP a conexões dedicadas, ou seja, a capacidade das portas disponíveis em seu *hardware* (servidor de comunicação). São as ligações mais procuradas por médias empresas que têm uma necessidade de tráfego maior. Começam com velocidades em 64 Kbps síncrono, sendo que estas conexões oferecem conectividade completa a Internet e todas as máquinas rodam TCP/IP estão visíveis na Internet. O custo de manutenção destas linhas não é baseado na quantidade de dados que são transmitidos, mas sim numa quantia fixa mensal estabelecida.

Segundo DODD (1998), ISDN é um padrão digital e público de enviar voz, vídeo, dados ou pacotes sobre a rede de telefonia pública e vem ganhando popularidade nos anos recentes devido a sua disponibilidade e velocidade. A tabela 5 apresenta uma comparação do tempo requerido para baixar um arquivo da Internet via modem e via ISDN. É importante salientar que o tempo despendido pelo modem depende também da qualidade da linha telefônica.

Tipo de Serviço	Tempo para Baixar um arquivo de 10 Mb
Modem 14.4 Kbps	93 minutos
Modem 28.8 Kbps	46 minutos
ISDN canal portador de 64 Kbps	21 minutos
ISDN canal portador de 128 Kbps	10 minutos

TABELA 5 – Comparação do tempo para baixar um arquivo de 10 Mb

3.4.3 Cable Modems

Utiliza as facilidades da TV a cabo para comunicação de dados. É um serviço semelhante a conexão dedicada, permitindo o acesso a Internet pelo mesmo mecanismo que leva às casas as imagens e os sons das televisões por assinatura. Os computadores ficam plugados o dia inteiro, sem custos adicionais, sem conta telefônica e com uma conexão de alta velocidade.

Segundo DODD (1998), a conveniência foi o motivo de criação do canal "inverso" para a comunicação de dados no mesmo cabo utilizado pela TV. A televisão é atualmente um meio de *broadcast* de uma via. Os sinas da televisão são transmitidos dos estúdios da TV, via satélite, através de microondas captadas pelas operadoras de TV a cabo. A partir destes, o sinal da TV é transmitido via cabo coaxial ou um canal híbrido (combinação de fibra ótica e cabo coaxial). A criação do canal inverso é feita utilizando diferentes freqüências para envio e recepção. O canal de envio a partir do emissor para a operadora de TV a cabo utiliza de 5 a 30 MHz ou 5 a 42 MHz. O canal de recepção utiliza 54 a 350 MHz ou 54 a 759 MHz. Dividindo a freqüência em duas taxas, possibilita que o cabo seja utilizado em diferentes velocidades. O mesmo cabo pode ser usado para enviar e receber. A figura abaixo ilustra como ocorre o entrelaçamento na casa do usuário.

Fonte: DODD, Annabel. *The Essential Guide to Telecommunications*.New Jersey: Prentice Hall PTR. 1998. 1ª Edição

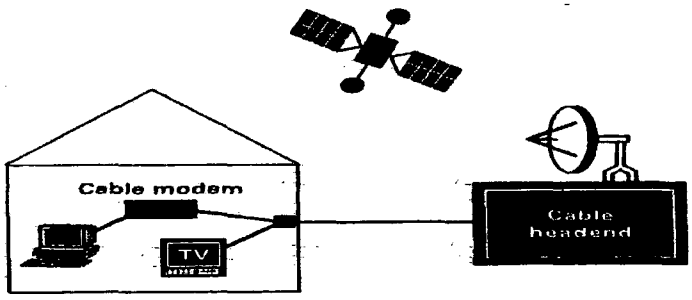


FIGURA 16 – Acesso via *cable modem*

Segundo LOPES (2000), há uma verdadeira explosão do mercado mundial de acesso de alta velocidade via cabo. Nos primeiros quatro meses de 2000, o segmento cresceu aproximadamente 70%. Os analistas avaliam que até o final deste ano o mercado de *cable modems* exceda 5,5 milhões de unidades vendidas, contra 2,8 milhões comercializadas em 99.

3.4.4 Rádio

A radiodifusão utiliza o sistema de redes sem fio (*wireless networks*) cujos pacotes são transmitidos através do "ar", em canais de freqüência de rádio (freqüências na faixa de KHz até GHz). É uma alternativa viável na qual é difícil ou mesmo impossível instalar cabos metálicos ou de fibra ótica. É muito utilizada em aplicações cuja confiabilidade do meio de transmissão é requisito indispensável. Nas ligações entre redes locais, a radiodifusão também tem papel relevante, especialmente se as redes estão distantes e o tráfego inter-redes é elevado. Nesse caso, circuitos telefônicos podem ser inadequados e a radiodifusão pode fornecer a largura de banda necessária. Quando se utiliza a radiodifusão como meio de transmissão, um aspecto que tem que ser levado em consideração é a segurança. Para garantir privacidade é indispensável a utilização de algum mecanismo de criptografia ao transmitir os sinais.

Através da utilização portadoras de rádio ou infravermelho, as WLANs estabelecem a comunicação de dados entre os pontos da rede. Os dados são modulados na portadora de rádio e transmitidos através de ondas eletromagnéticas. Múltiplas portadoras de rádio podem coexistir sem que uma interfira na outra. Para extrair os dados o receptor sintoniza numa freqüência específica e rejeita as outras portadoras de freqüências diferentes.

Vários artigos relacionados a este assunto vêm sendo publicados. Dentre os quais pode-se citar o artigo de LIU & MORDOWITZ (1998), que comprovou, através de simulação, que com a utilização de *wireless* pode-se ampliar a complexidade da rede, ou seja, aumentar o número de computadores/nodos e ainda assim obter um ganho de

tempo e de eficiência. O ganho de tempo é de 108% para 164% e de eficiência de 54% para 82%, quando se aumenta o número de nodos de 20 para 160.

Num ambiente típico, como o mostrado na Figura 17, o dispositivo transceptor (transmissor/receptor) ou ponto de acesso (*access point*) é conectado a uma rede local Ethernet convencional (com fio). Os pontos de acesso não apenas fornecem a comunicação com a rede convencional, como também intermediam o tráfego com os pontos de acesso vizinhos, num esquema de micro células com *roaming* semelhante a um sistema de telefonia celular.

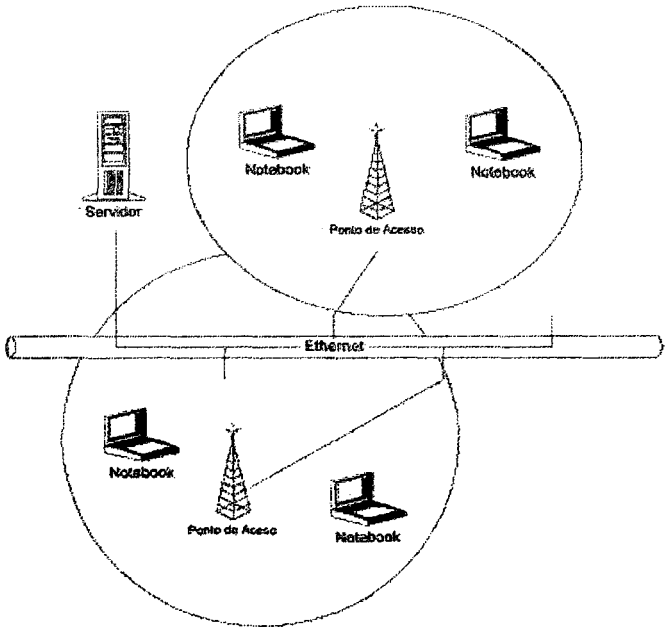


FIGURA 17 – Acesso via rádio

3.5 Considerações Finais

Este capítulo descreveu as características, funções, arquitetura e equipamentos utilizados no ambiente de provimento de serviços Internet com o objetivo de servir como base para as demais etapas da dissertação.O capítulo que segue apresenta uma aplicação da metodologia de planejamento de capacidade através de um estudo de caso em um ambiente de provimento de serviços de pequeno porte.

Capítulo 4

Estudo de Caso

Este capítulo apresenta as primeiras etapas da aplicação da proposta de adequação da metodologia de planejamento de capacidade no ambiente de provimentos de serviços Internet. Para tanto, está dividido em quatro seções. A primeira descreve o ambiente no qual o trabalho é realizado. A segunda fala sobre a definição dos objetivos de negócios. A terceira relata o processo de caracterização da carga de trabalho e a última seção apresenta uma previsão desta carga.

4.1 Compreensão do ambiente

Conforme descrito no capítulo 2 a compreensão do ambiente consiste no estudo do *hardware*, do *software*, dos elementos de conectividade, protocolos, períodos de pico e os níveis de qualidade de serviços definidos. Para isso, MENASCÉ & ALMEIDA (1999) colocam algumas questões que auxiliam na realização desta etapa:

- Qual a configuração dos servidores?
- Qual a configuração da rede?
- Qual a capacidade da linha de comunicação de dados?
- Qual sistema operacional está sendo utilizado?
- Qual é a descrição dos demais componentes, como hubs, roteadores, *firewall*, *servidor proxy*?

Para a efetiva aplicação do método foram contatadas algumas empresas do setor de provimento de serviços Internet para verificar a possibilidade de realização de uma parceria. Depois de escolhida uma delas, foram realizadas várias reuniões com o proprietário com vistas à obtenção dos dados abaixo.

4.1.1 Clientes e meios de conexão

O provedor de serviços adota uma política de diversificação de seus clientes com o objetivo de otimizar a sua capacidade instalada. Para tanto, trabalha com o segmento de pessoas físicas e jurídicas (a utilização dos serviços por estes clientes restringe-se basicamente ao horário comercial). Os seus clientes e suas formas de acesso são:

Clientes	**Qtde**	**Meio de Acesso**
Pessoas Físicas	60	Conexão por discagem
Pessoas Físicas – Condomínios	150	Rádio
Pessoas Jurídicas	30	Conexão por discagem
Pessoas Jurídicas	1 empresa(135 máq)	Conexão dedicada

TABELA 6 – Tipo de clientes e conexões

4.1.2 Equipamentos

O provedor de serviços é de pequeno porte e seus equipamentos são:

- 20 Linhas telefônicas;
- 20 Modems US Robotics de 56K - X2;
- 01 Roteador CISCO 2511;
- 01 Roteador CISCO 2522;
- 01 Link de 512 Kbps com a empresa Brasil Telecom;
- 01 Hub IBM;
- 01 servidor 486 DX 266, com 12 Mb de memória RAM, HD IDE de 127 Mb, com velocidade de 2 Mb/s, sistema operacional Windows NT, exercendo função de *gateway* do acesso de rádio;
- 01 servidor Pentium 200, com 64 Mb de memória RAM, HD Scsi com capacidade para 4 Gb e velocidade de 10 Mb/s, sistema operacional LINUX, exercendo a função DNS;

- 01 servidor Pentium, com 24 Mb de memória RAM, HD IDE com capacidade de 730 Mb e velocidade de 3 Mb/s, sistema operacional Windows NT, exercendo a função de atendimento ao acesso discado;
- 01 servidor K6 II 266, com 64 Mb de memória RAM, HD IDE de 8 Gb com velocidade de 6 Mb/s, sistema operacional LINUX, exercendo a função de proxy. O proxy utilizado é o software SQUID;
- Rede interna: Ethernet CSMA/CD.

4.1.3 Acesso via sistema de rádio

A empresa oferece acesso através de rádio para moradores de condomínios e empresas, provendo flexibilidade e mobilidade. Para tanto, uma antena foi implantada no Morro da Cruz e, quando da instalação de um acesso num determinado local, é providenciada a colocação de uma antena e um servidor no topo de cada edifício(exerce a função de *gateway*). Em função da segurança, cada servidor destes tem um endereço *MAC Adress Filtering*, e o *Wave Point* (servidor instalado no Morro da Cruz) só responde se este endereço estiver na sua lista e se tiver o nome correto da rede. Umas das características mais importantes com relação ao desempenho deste sistema é o alto *throughput*. Hoje as redes *WaveLAN* IEEE rodam a taxa de 02 Mb.

4.1.4 Acesso via conexão dedicada

A empresa oferece acesso aos seus serviços para uma empresa aqui chamada de Empresa 1 através de uma conexão dedicada de 64 kb.

4.1.5 Nível de qualidade de serviços

A empresa não estabeleceu com os seus clientes um nível de qualidade de serviços.

4.1.6 Horários de pico

Como pode ser observado nos gráficos a seguir, que representam a utilização do *link* de comunicação com a Brasil Telecom, o canal apresenta-se quase que totalmente utilizado no período compreendido entre as 08 e 02 horas. Estes gráficos foram obtidos através da ferramenta de monitoração de tráfego NRTJ.

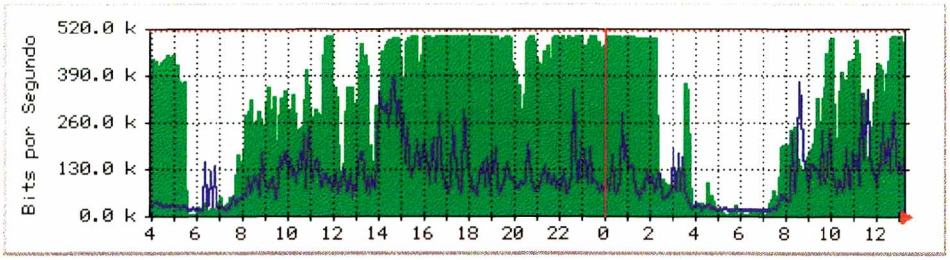


FIGURA 18 – Gráfico Tráfego Diário (14/09/2000)

Max In: 501.9 kb/s (98.0%) Média In: 314.8 kb/s (61.5%) Corrente In: 483.2 kb/s (94.4%)
Max Out: 382.4 kb/s (74.7%) Média Out: 106.7 kb/s (20.8%) Corrente Out: 175.4 kb/s (34.3%)

TABELA 7 – Tráfego diário

Legenda:

VERDE	Tráfego entrando no provedor, em bits por segundo.
AZUL	Tráfego saindo do provedor, em bits por segundo.

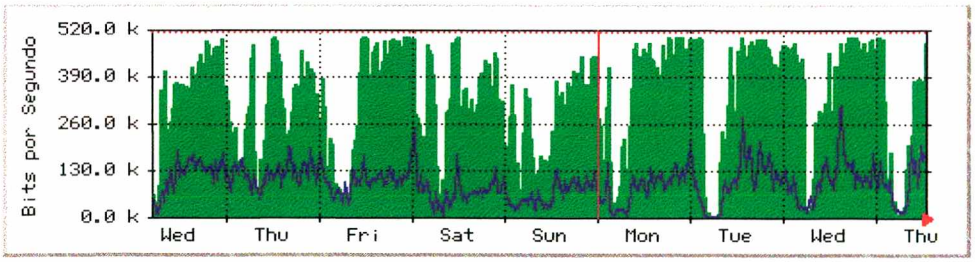


FIGURA 19 – Gráfico Tráfego Semanal (08/09/2000 a 14/09/2000)

Max In: 501.1 kb/s (97.9%) Média In: 328.1 kb/s (64.1%) Corrente In: 488.7 kb/s (95.5%)
Max Out: 309.1 kb/s (60.4%) Média Out: 96.0 kb/s (18.8%) Corrente Out: 181. kb/s (35.4%)

TABELA 8 – Tráfego Semanal

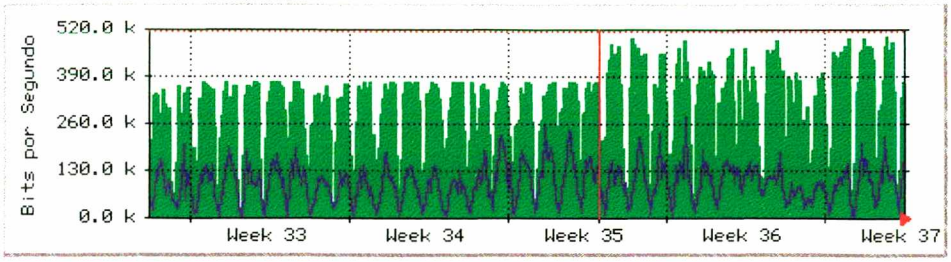


FIGURA 20 – Gráfico Tráfego Mensal (15/08/2000 a 14/09/2000)

Max In:	500.7 kb/s (97.8%)	Media In:	301.1 kb/s (58.8%)	Corrente In:	376.4 kb/s (73.5%)
Max Out:	276.8 kb/s (54.1%)	Média Out:	95.5 kb/s (18.7%)	Corrente Out:	180.2 kb/s (35.2%)

TABELA 9 – Tráfego Mensal

Analisando o gráfico semanal pode-se verificar que o padrão do gráfico diário é praticamente o mesmo para todos os dias da semana, não existindo sazonalidade semanal evidente.

4.1.7 Proxy

O *software* utilizado para gerenciar o *proxy* é o Squid. É um servidor de alto desempenho para clientes web que suporta os protocolos FTP, Gopher e HTTP.

No provedor são armazenados somente os arquivos do tipo HTML. A política de substituição de arquivos utilizada é o LFU sendo excluídos os arquivos menos acessados na semana.

4.1.8 Topologia

Um desenho da topologia do ambiente é apresentado na fig. 21.

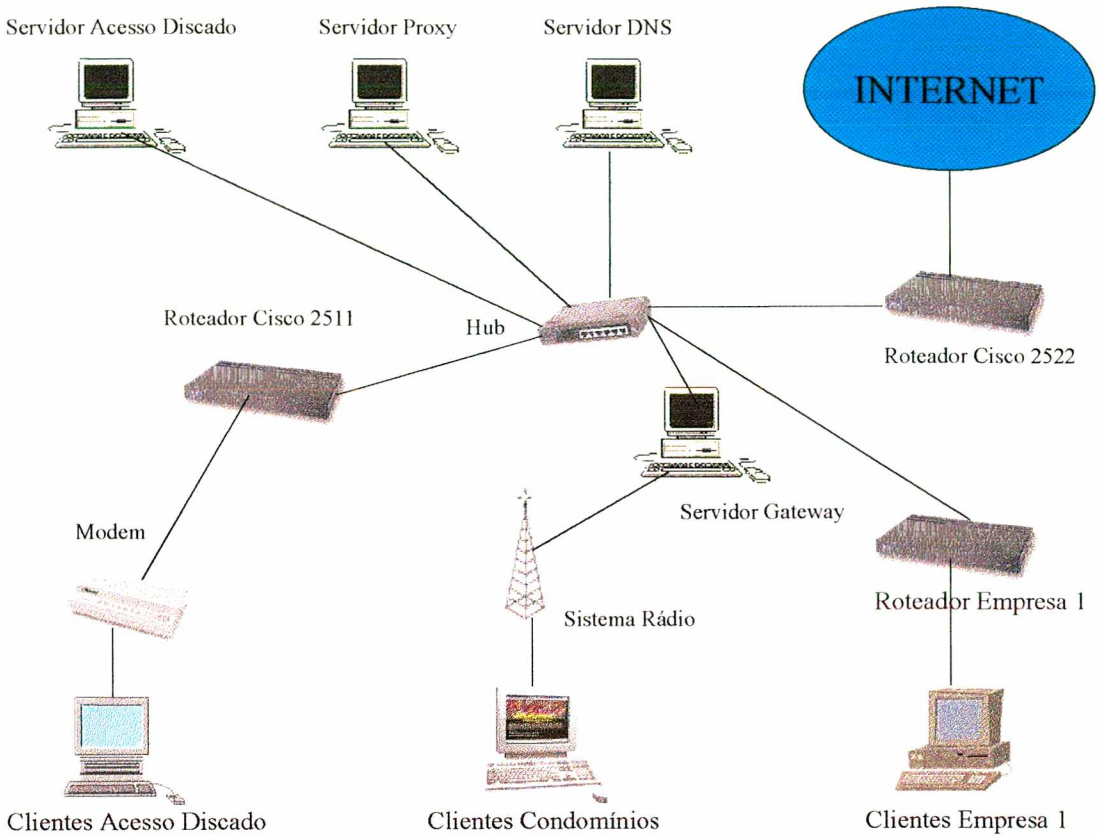


FIGURA 21 – Topologia do provedor de serviços Internet

4.2 Definição dos objetivos de negócios

O objetivo da empresa é concentrar e aumentar sua atividade, atendendo ao segmento de condomínios através do acesso via rádio. A meta é aumentar em média 15 clientes por mês. Justifica esta escolha pelo retorno sobre o investimento deste segmento e pela qualidade dos serviços que pode ser oferecida através desta forma de acesso. Esta tendência já pode ser observada no fig. 22, em que se verifica a diminuição gradual dos clientes de acesso discado e o aumento de clientes utilizando a conexão via rádio.

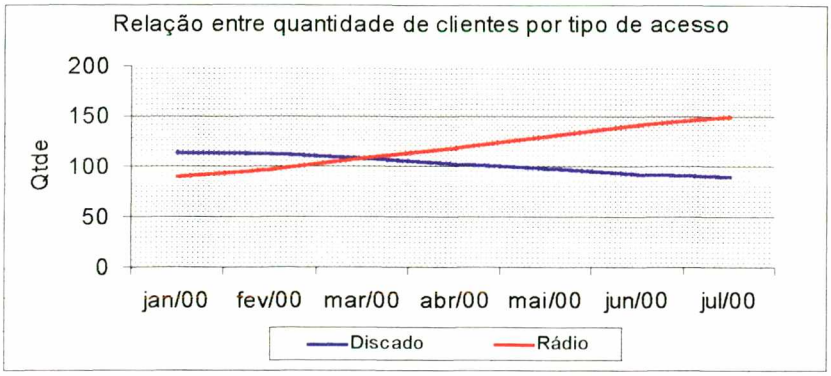


FIGURA 22 - Tendência do crescimento dos clientes

4.3 Caracterização da carga de trabalho

A metodologia proposta foi aplicada no ambiente web e no que tange ao planejamento de capacidade é importante ressaltar as peculiaridades deste, em relação aos ambientes tradicionais. Apresenta características únicas, principalmente no processo de caracterização da carga, que exercem um profundo impacto no desempenho geral e que devem ser tratadas com cuidado especial.

Uma destas características é a imprevisibilidade do comportamento dos usuários. Por ser um canal relativamente novo e em constante desenvolvimento, é difícil saber com antecedência como e quando os usuários irão utilizar os serviços, ou seja, estes acessos são completamente aleatórios. Outra questão é o crescimento da quantidade do número de usuários. Além disso deve ser considerado o tipo de conteúdo. Segundo o consultor da Ernst & Young, Klauber Santos em LOPES (2000b): "A necessidade de trafegar som, imagem, vídeo e outros objetos de dezenas de Megabytes de tamanho, sem uma idéia clara do volume e da freqüência que isso estará ocorrendo, dificulta bastante os trabalhos de dimensionamento".

A web também é caracterizada pela larga diversidade de componentes, que complicam o monitoramento e coleta de dados. São diferentes *browsers* e servidores executando em uma variedade de plataformas com diferentes capacidades. Isto faz com que em certas ocasiões, os usuários da web podem ter de esperar por um *delay* longo, variável e imprevisível, na espera pela realização de suas tarefas, dependendo da largura de banda e congestionamento da rede.

Todas estas características dificultam a previsão de carga de trabalho e fazem com que sejam necessárias algumas simplificações no processo de caracterização de carga neste ambiente.

4.3.1 Busca de parâmetros para caracterização da carga

Como já citado no capítulo 1, somente foi disponibilizada uma pequena amostra do ambiente em estudo e esta foi complementada com *log* de dados de um ambiente semelhante. Não foi possível extrair todos os dados necessários para a caracterização da carga, no entanto esta é uma área de grande interesse de vários pesquisadores e muitas são as publicações que versam sobre o tema. Optou-se então por verificar a caracterização da carga apresentada nas publicações e complementá-la com dados obtidos através da análise dos *logs*. Para tanto um pequeno relato das principais publicações nesta área é descrito a seguir.

ARLIT & WILLIAMSON (1997a) explicam que a web é baseada no modelo cliente-servidor. A comunicação é sempre na forma do par requisição-resposta e é sempre iniciada pelo cliente. Este, acessando documentos na web, utiliza um *browser* que envia as requisições para serem respondidas pelo servidor web. Cada página web pode consistir em múltiplos documentos e cada um destes arquivos é requisitado separadamente para o servidor. Este, por sua vez, interpreta o pedido do cliente e responde às requisições. Não se pode desconsiderar o *overhead* existente que, conforme ZWIEBACK, é de aproximadamente 30% para constantemente abrir e fechar conexões e segundo MENASCÉ & ALMEIDA (1999) é em torno de 20%. YU (1998) coloca que

a resposta para uma requisição pode ser lenta, especialmente se os servidores estão longe do cliente, ou conectados através de uma conexão lenta ou *link* ocupado.

Em MENASCÉ & ALMEIDA (1999) trata-se exatamente das características que distinguem o comportamento do ambiente web dos ambientes tradicionais. Uma delas é a extrema variabilidade da carga de trabalho. Por exemplo, através da análise de *web sites* verifica-se que poucos documentos têm um tamanho menor que 100 Bytes. A maioria tem em média centenas de Bytes até Megabytes. CROVELLA & BESTAVROS (1997) comprovaram que os arquivos variam numa amplitude de 100 Bytes até 10 Mb.

ARLIT & WILLIAMSON (1997a) em seu artigo afirmam que encontrar uma distribuição dos tamanhos dos arquivos que transitam pela web é uma tarefa difícil. Os tipos de documentos acessados pelos clientes podem ser classificados em inúmeras categorias. O artigo menciona que os estudos sobre a carga de trabalho na web indicam que mais de 90% das requisições dos clientes são referentes ao protocolo HTTP; dentro destes, a maioria pequenos arquivos HTML ou imagens. MAHARTI & WILLIANSON (1999) afirmam que 95% das requisições são documentos HTML e imagens. MENASCÉ & PERAINO (1999), por sua vez, afirmam que as requisições do protocolo HTTP representam 99,72% de toda a carga de trabalho que transita pela web.

ARLIT & WILLIAMSON (1997a), CROVELLA & BESTAVROS (1997) e MAHARTI & WILLIANSON (1999) sugerem que a distribuição dos tamanhos dos arquivos segue uma distribuição de Pareto. Esta distribuição tem a característica de ter a cauda da distribuição declinando lentamente. Isto significa que quando variáveis randômicas são geradas a partir de uma distribuição deste tipo, a probabilidade de obter valores extremamente grandes é não-negligenciável (i.e. Pareto, com α.1,30). A distribuição de tamanhos dos documentos na web significa que tem arquivos muito grandes (*outliers*) na cauda da distribuição que são relativamente poucos em quantidade, mas são grandes o suficiente para contribuir no aumento substancial do volume observado.

A definição matemática desta distribuição é:

$$P[X > x] \sim x^{-\alpha}, \quad x \rightarrow \infty, \quad 0 < \alpha < 2 \qquad (4.1)$$

ARLIT *et al* (1999) relatam as medições realizadas em um *proxy* de um provedor de serviços Internet que fornece acesso através de um cabo de alta velocidade. Este artigo enfoca as características da carga de trabalho como distribuição por tipo de arquivo, por tamanho de arquivo e por comportamento dos arquivos referenciados. Apresentam-se as seguintes conclusões:

- Requisições HTTP são em torno de 99,3% e representam 87,7% do tráfego, sendo que destes arquivos, aproximadamente 73%, são de imagens e em torno de 12% são arquivos HTML. Arquivos de imagens e arquivos HTML respondem por apenas um pouco mais da metade dos dados transferidos. Isto é significantemente pouco comparado com o percentual de requisições realizadas. Os arquivos pesados são os de áudio, vídeo, compactados e executáveis. Enquanto estes arquivos representam apenas 1,1% das requisições, são responsáveis por 37,8% do tráfico;

- O outro protocolo significante em termos de quantidade de dados distribuídos é o FTP que solicita 0,3% das requisições, entretanto representa 12,1% dos dados;
- Os clientes ficam mais propensos a solicitar arquivos maiores quando a velocidade da rede e tempo de resposta aumentam;

- O uso dos serviços cresce em função de novos usuários e do aumento de utilização dos serviços dos usuários atuais;

- O padrão dos arquivos requisitados não é uniforme;

- Foram encontrados poucos ganhos ao se fazer *cache* de arquivos FTP pela pouca quantidade de requisições destes tipos de arquivo e por consumirem muito espaço de *cache*.

4.3.1.1 Obtenção da carga de trabalho

Para encontrar as características da carga de trabalho é necessário utilizar um *log* de acesso. No ambiente de provimento de serviços Internet o ideal é utilizar o *log* do servidor *proxy*. ARLIT *et al* (1999) descreve como obter estes dados, através da análise de cada entrada no *log* de acesso no servidor web, que contém informações sobre requisições simples de cada cliente. Cada entrada inclui as seguintes informações:

- Endereço do cliente: endereço dinâmico do cliente associado após a conexão ao provedor de serviços;
- Data e hora que a requisição foi feita;
- R*equest*: contém o protocolo utilizado pelo cliente;
- Código de status: indica a natureza da resposta;
- C*ontent* dados: quantidade de dados em bytes passados entre o cliente e o servidor;
- Tempo de transferência: a quantidade de tempo entre a chegada da requisição do cliente e resposta do *proxy* ou servidor;

Algumas das informações não podem ser examinadas por que não são registradas no *log*. Por exemplo, quando o usuário faz uma requisição e fica impaciente com a demora e aborta a requisição antes da mesma ser completada. Estima-se que em torno de 10,3% das requisições sejam abortadas, e segundo FELDMANN *et al* (1999) num provedor de acesso *dial up* apresenta-se um número similar.

4.3.1.2 Modelos de carga

Vários pesquisadores apresentaram modelos de carga de trabalho por tipo de arquivo e freqüência de solicitação. Apresentam-se a seguir os principais.

Segundo ARLIT *et al* (1999) a carga é caracterizada em termos de protocolo, conforme a tabela 10:

Item	HTTP	FTP	Gopher	Outros
% Requisições	99,30	0,30	0,02	0,38
% Dados	87,70	12,10	0,03	0,17
Tamanho médio (Kb)	10,6	432	14,4	5,7

TABELA 10 – Requisição por tipo de protocolo

Em termos de protocolo HTTP a carga é caracterizada conforme segue:

Item	HTML	Imagens	Áudio	Vídeo	Format	Compact	Exe	Outros
% Req	12,4	73,1	0,6	0,2	0,00	0,2	0,1	13,4
% dados	4,8	47,6	3,9	19,9	0,2	5,8	8,3	9,5
Média(byes)	6,354	14,032	135,734	1,593,565	247,374	553,781	1,642,792	25,856
Mediana(bytes)	3,051	4,694	37,806	925,735	79,920	92,263	766,692	5,719

TABELA 11 – Requisição do protocolo HTTP

MENASCÉ & ALMEIDA (1999) definiram o modelo de carga de trabalho de um site padrão conforme abaixo:

Tipo requisição	Tamanho médio (Kb)
HTML pequeno	2,0
HTML grande	10
Pesquisa CGI	2,5
Imagem	35
Som	120
CGI script	1,5

TABELA 12 – Requisição do protocolo HTTP

YU (1998) através da análise das requisições HTTP obteve a seguinte caracterização da carga:

Tipo	Quantidade	**Freqüência %**
Arquivos aplicação	6	0,001305
Áudio	104	0,022617
Imagem	307504	66,873106
Texto	99231	21,579834
Vídeo	106	0,023052
Erro	52881	11,500067

TABELA 13 - Tipo dos arquivos e distribuições por requisição

4.3.1.3 Comportamento dos usuários

Uma importante observação é feita por ARLIT *et al* (1999) com relação ao uso dos serviços web. Foi comprovado que a carga de trabalho é afetada pela rotina dos usuários, ou seja, o crescimento da carga não ocorre somente pelo aumento de novos usuários, mas também devido ao aumento da freqüência de utilização dos que já são clientes, que passam a se acostumar com os serviços e passam a utilizá-lo com mais intensidade.

Outra consideração interessante é que o padrão de acesso dos usuários muda quando a conexão com a Internet não tem gargalos. Os usuários ficam mais inclinados a solicitar arquivos maiores quando a largura de banda aumenta substancialmente. Este comportamento vai gerar cada vez mais carga nos servidores e no *link* disponível.

O comportamento dos usuários com acesso veloz (*cable modems*, ISDN) é bem diferente daqueles com acesso mais lento, como *dial up*. Características como tamanho dos arquivos e freqüência entre chegadas vão variar bastante.

4.3.2 Caracterização da carga no ambiente em estudo

Para a efetiva caracterização do ambiente em estudo procurou-se seguir os passos gerais identificados no capítulo 2, item 2.1.2.2.1. Fazendo um Esboço do Modelo

verifica-se que os componentes básicos do ambiente em estudo são formados pelas requisições dos usuários e as respectivas respostas dos servidores. As primeiras compreendem:

- Tipo de requisição;
- Tamanhos dos arquivos;
- Taxa de chegada (freqüência com que as requisições chegam ao provedor).

Com relação às respostas do servidor deverão ser considerados:

- Tipo do arquivo;
- Tamanho do arquivo de resposta: varia muito principalmente em função do tipo de protocolo, podendo ser um simples arquivo de texto, ou então arquivos mais pesados com imagens, som e vídeo;
- Tempo de resposta: depende de vários fatores, muitos deles incontroláveis, como congestionamento em *links*, congestionamento em servidores, gargalo de roteadores, tamanho do arquivo de resposta.

Para a realização da etapa de coleta de dados foram encontradas algumas dificuldades porque a empresa somente disponibilizou uma amostra inexpressiva das requisições dos usuários. Para complementar os estudos foram analisadas 600 mil linhas de requisições de usuários no *proxy* Squid, obtidas a partir de um ambiente semelhante, ou seja, um acesso dedicado. A partir da amostra foi possível extrair o tamanho médio e a freqüência de chegada por tipo de arquivo e verificar a taxa de chegada das requisições por usuário. Estes valores são validados por ocasião da utilização do modelo de carga no modelo de desempenho e comparação com o percentual de utilização dos recursos do ambiente real utilizando a carga atual.

É sabido que a cada dia surgem novos serviços disponibilizados no ambiente web, que vão sendo solicitados pelos usuários à medida de suas necessidades e curiosidade. A partir desta constatação e sabendo da diversidade do comportamento dos usuários uma caracterização mais precisa por tipo de arquivo requisitado necessita de uma análise apurada dos *logs* e aplicação das técnicas para a caracterização da carga. Como os *logs* disponibilizados permitiram identificar somente o protocolo HTTP, a alternativa foi

complementar estes dados com os artigos descritos acima. Analisando os vários modelos de carga foi tomado como base a caracterização proposta por ARLIT *et al* (1999), pois a mesma é atual, detalhada e foi obtida num ambiente de conexão dedicada, semelhante ao ambiente em estudo. As seções a seguir descrevem os parâmetros utilizados como modelo de carga.

4.3.2.1 Modelo de Carga - requisições dos clientes

Com relação às requisições dos clientes os parâmetros utilizados são:

- **Tipo de requisição:**
  - Protocolo FTP;
  - Outros Protocolos;
  - Protocolo HTTP (HTML, Imagens, Vídeo, Áudio, Compactados, Executáveis, Outros);
- **Tamanhos dos arquivos**: conforme MENASCE & ALMEIDA (2000), o tamanho médio de uma requisição neste ambiente é de 200 bytes;
- **Taxa de chegadas**: a tabela 14 apresenta a freqüência entre requisições obtida a partir dos artigos e a tabela 15 apresenta a freqüência obtida através da análise dos *logs*. Pode-se observar a similaridade entre as duas tabelas. A tabela 15 não apresenta os demais protolocos. Para o modelo de carga é utilizado uma média das duas tabelas

Protocolo	Tipo	% Req
FTP		0,30
Outros Protocolos		0,40
HTTP	HTML	12,31
	Imagens	72,59
	Audio	0,60
	Video	0,20
	Compact	0,20
	Executável	0,10
	Outros	13,31
	Sub total	99,30
Total		**100,00**

TABELA 14 – Freqüência de chegada por tipo de arquivo - obtida nas referências

Protocolo	Tipo	% Req
HTTP	HTML	16,11
	Imagens	73,89
	Audio	0,06
	Video	0,02
	Compact	0,13
	Executável	2,73
	Outros	7,06
TOTAL		100,00

TABELA 15 - Freqüência de chegada por tipo de arquivo – obtida na análise *log*

Protocolo	Tipo	% Req
FTP		0,30
Outros Protocolos		0,40
HTTP	HTML	14,16
	Imagens	72,98
	Audio	0,33
	Video	0,11
	Compact	0,16
	Executável	1,41
	Outros	10,16
SUBTOTAL		99,30
TOTAL		**100,00**

TABELA 16 - Freqüência de chegada por tipo de arquivo - média

A partir da freqüência por tipo de arquivo faz-se necessário ainda determinar quantas requisições um usuário solicita normalmente num determinado período. Esta quantidade de requisições tem reflexo direto na intensidade da carga. Para tanto, o *log* foi classificado por usuário e analisado o seu comportamento, ou seja, foi verificado quantas requisições uma pessoa realiza em média num determinado período. A quantidade obtida foi de 630 arquivos em uma hora. Para utilização no modelo o período de tempo utilizado é segundos. Portanto foi realizada a seguinte dedução (exemplo: arquivos de Imagem):

- Em 1 hora são feitas 630 requisições/usuário;
- Do total das 630 requisições, 72,98% são requisições de Imagens, o que corresponde a 459,77 requisições (250 * 72,98%= 459,77);

- Distribuindo no período de 3600 segundos observa-se que a cada 8 segundos chega uma requisição do tipo Imagem no Provedor de Serviços/usuário (3600/459,77=8).

Este cálculo foi realizado para todos os tipos de arquivos, obtendo-se a seguinte taxa de chegada:

Protocolo	Tipo	Taxa Chegada (seg)
FTP		1.905
Outros Protocolos		1.428
HTTP	HTML	40
	Imagens	8
	Audio	1.744
	Video	5.231
	Compact	3.487
	Executável	406
	Outros	56

TABELA 17 – Taxa de chegada por tipo de arquivo

Contudo as taxas entre chegadas relacionadas na tabela 17 têm um comportamento determinístico e sabe-se que num ambiente real isto é praticamente impossível. Partiu-se então para a busca de uma distribuição que represente melhor o comportamento de um usuário padrão. O *log* foi então classificado por usuário, tipo de requisição e horário. Neste ponto, utilizou-se a ferramenta Input Analizer do Software de Simulação de Sistemas - ARENA, que realizou a função de verificar a melhor distribuição dada a taxa entre as chegadas por usuário e tipo de requisição.

Como pode ser visualizado nas figuras abaixo, a distribuição que melhor representa esta taxa é a Exponencial. Os valores médios obtidos na análise de vários usuários foram os relacionados na tabela 18.

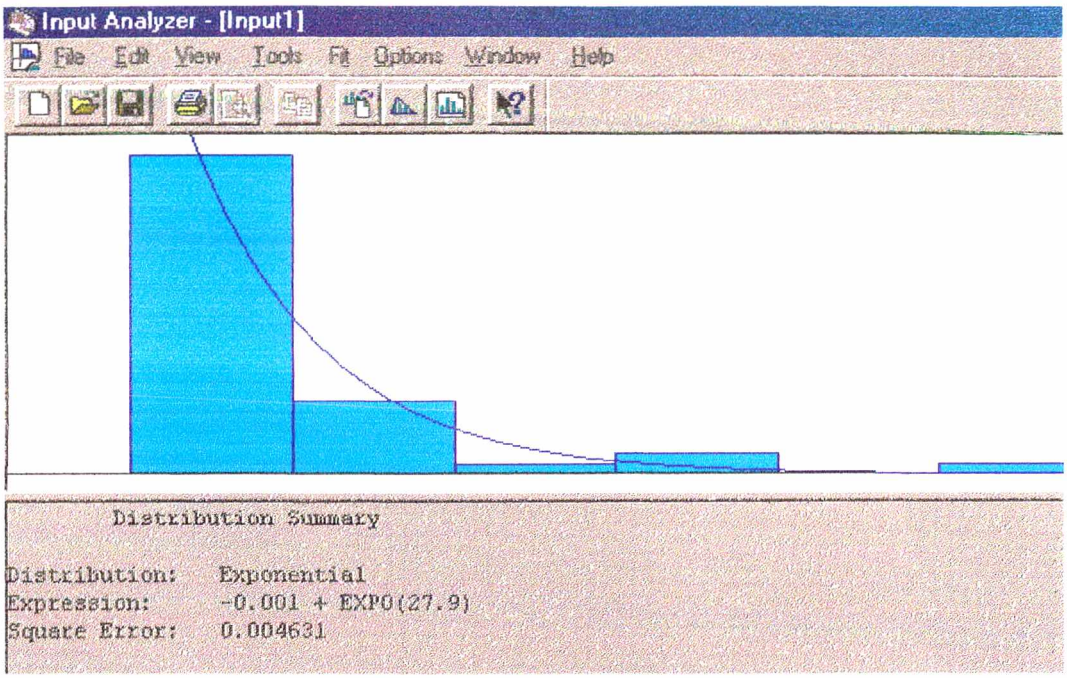


FIGURA 23 - Arquivos HTML

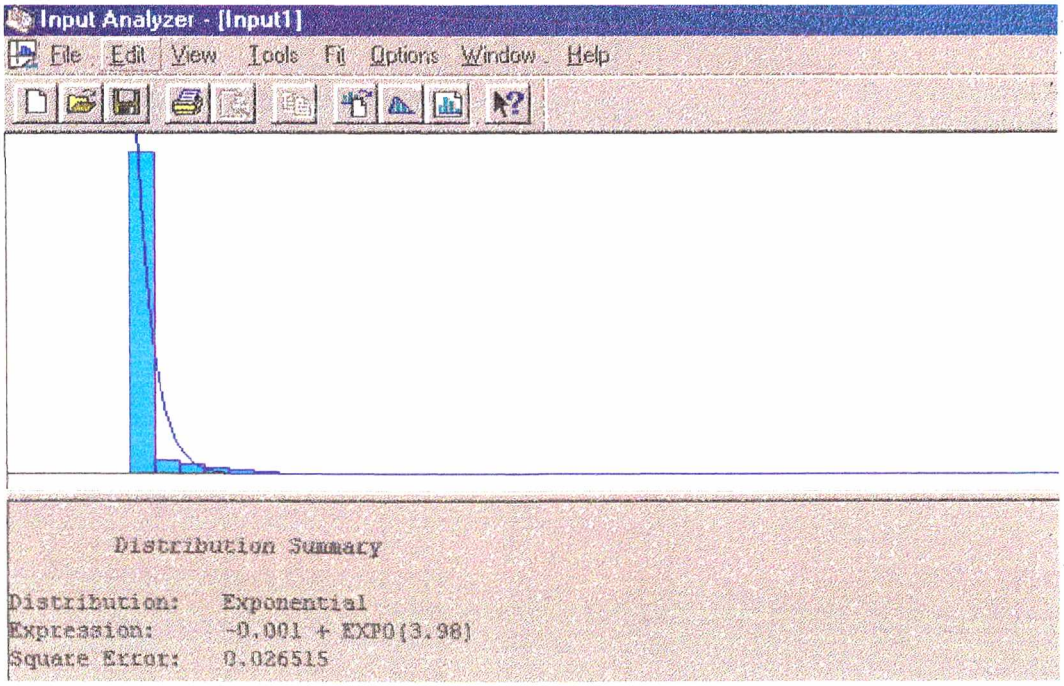


FIGURA 24 - Arquivos de imagens

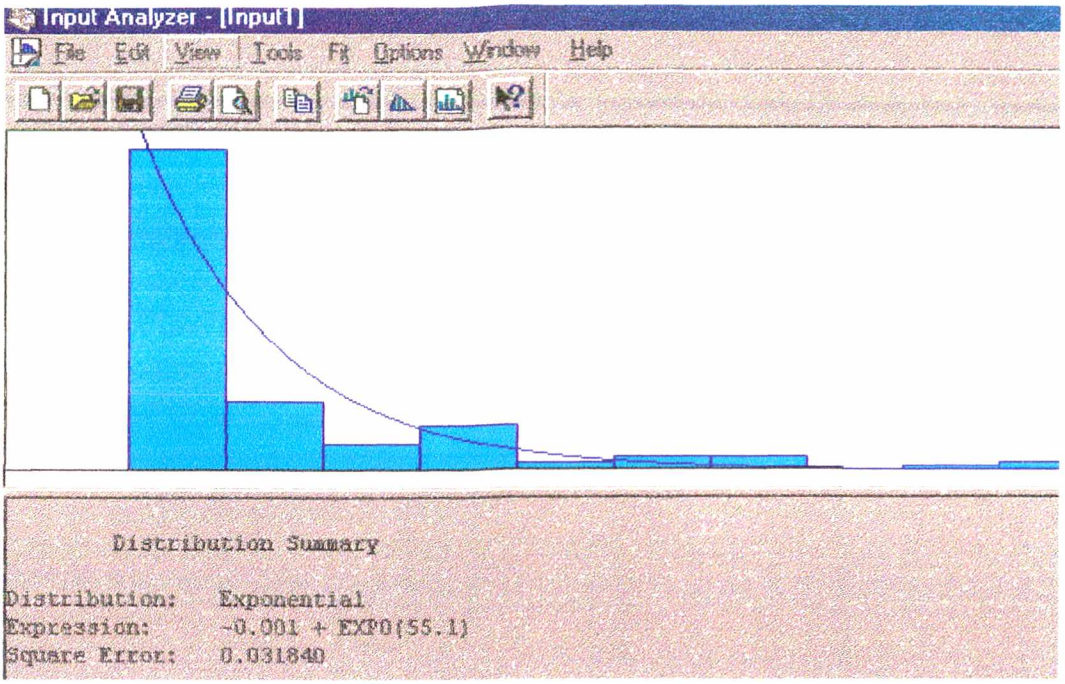


FIGURA 25 - Arquivos - Outros

Protocolo	Tipo	Taxa Chegada (seg)
HTTP	HTML	29,25
	Imagens	5,46
	Outros	45,25

TABELA 18 – Taxa de chegada por tipo de arquivo obtido a partir do Input Analizer

Não foi possível analisar todos os tipos de arquivos listados na tabela 17. Decidiu-se utilizar como taxa entre chegadas a distribuição Exponencial e os valores médios relacionados na tabela 17.

4.3.2.2 Modelo de Carga – respostas dos servidores

Com relação às respostas dos servidores obteve-se o seguinte modelo de carga:

- **Tipo do arquivo:** depende do tipo de requisição efetuada pelo usuário;

- **Tamanho do arquivo de resposta**: A tabela 19 representa os valores obtidos através dos artigos. O tamanho médio dos Protocolos FTP e Outros Protocolos obtido da tabela 10 foi multiplicado por 1024, pois a tabela apresenta os valores em Kb e no modelo é utilizado *bytes* como medida padrão. Os valores referentes ao protocolo HTTP foram obtidos a partir da tabela 11. A Tabela 20 apresenta os tamanhos médios obtidos através da análise do *log* do *proxy*.

Protocolo	Tipo	Tamanho Médio (Bytes)
FTP		442.368
Outros Protocolos		20.582
HTTP	HTML	6.354
	Imagens	14.032
	Audio	135.734
	Video	1.593.565
	Compact	553.781
	Executável	1.642.792
	Outros	25.856

TABELA 19 – Tamanho médio por tipo de arquivo obtidos nas referências

Protocolo	Tipo	Tamanho Médio
HTTP	HTML	7.389
	Imagens	3.317
	Audio	100.251
	Video	215.224
	Compact	545.167
	Executável	46.602
	Outros	5.942

TABELA 20 – Tamanho médio por tipo de arquivo obtido através da análise do log

Analisando as duas tabelas anteriores verifica-se um tamanho menor nos arquivos da tabela 20. A primeira carga foi analisada em 1999 e a segunda em julho de 2000. Com a velocidade das mudanças no ambiente web as pesquisas tem evoluído no sentido de buscar maior benefício com menor custo. O mesmo é válido para o tamanho dos arquivos, ou seja, procura-se passar mais informação com arquivos cada vez menores, que utilizem menos recursos. Portanto, o tamanho médio dos arquivos definido para o Modelo de Carga são os da tabela 20, complementados com a tabela 19 nas informações faltantes. Como citado em vários artigos a distribuição dos tamanhos dos arquivos é

heavy-tailed (i.e. Pareto, com α.1,30). Na modelagem é utilizada esta distribuição, com o valor médio dos arquivos, conforme tabela 21.

Sabe-se que os tipos de arquivos requisitados pelos clientes de um provedor de serviços são em número muito maior dos que os descritos na tabela acima. Entretanto, em função das dificuldades surgidas na obtenção da carga e das limitações da ferramenta de modelagem, serão utilizados estes tipos de arquivos como medidas representativas.

4.4 Previsão da carga de trabalho

Previsão da carga de trabalho é o processo de predizer como a carga de trabalho do sistema irá se comportar no futuro. Isto envolve avaliar a tendência dos dados históricos e avaliar/analisar os negócios e planos estratégicos da organização e mapear estes planos para trocas no processo dos negócios. Segundo MENASCÉ & ALMEIDA (2000), durante o processo de previsão da carga, os componentes básicos da carga são associados ao processo de negócio e a troca na intensidade da carga destes componentes pode ser derivada do processo do negócio e dos planos estratégicos.

Métodos de previsão podem ser qualitativos e quantitativos. O modelo pode contar com dados históricos existentes para estimar valores futuros dos parâmetros da carga de trabalho.Contudo no ambiente web estes dados históricos não assumem tanta importância, tendo em vista a imprevisibilidade, peculiar a este ambiente. A aproximação qualitativa é mais aplicável, baseado em julgamentos intuições, opiniões de experientes, analogia histórica, conhecimento comercial e qualquer outra informação relevante.

Isto posto e analisando a fig. 22 que mostra a tendência de crescimento do número de clientes, seguindo o planejamento dos negócios da empresa, conclui-se que o aumento na carga do sistema em estudo será causada pelo aumento da quantidade de clientes de condomínios utilizando os serviços. O aumento da quantidade destes clientes

implica uma intensidade maior de requisições, aumentando a carga de trabalho, que é gerada sobre os componentes do ambiente.

4.5 Considerações finais

Muitas pesquisas têm sido feitas para melhorar o desempenho e escalabilidade da web. Os fatores chave para otimizar o desempenho são aqueles que irão reduzir o tráfego na rede produzido por clientes e servidores e melhorar o tempo de resposta para os usuários. Para isso faz-se necessário conhecer bem o ambiente e a carga de trabalho utilizada neste meio.

Este capítulo descreveu estas fases, com o objetivo de embasar as demais etapas da metodologia que estão descritas no capítulo seguinte, em que é feita a aplicação de um projeto experimental e a conseqüente avaliação de desempenho dos possíveis cenários.

Capítulo 5

Experimentação e análise dos resultados

Uma disciplina de gerenciamento muitas vezes encarada como gasto, o planejamento de capacidade significa na maioria dos casos, a otimização dos recursos da rede, dificultando o surgimento de falhas que tiram os sistemas do ar, além de garantir que a organização não esta jogando dinheiro fora na aquisição de máquinas ou *softwares* em excesso ou errado.

Este capítulo exemplifica a necessidade e utilidade da realização do planejamento de capacidade na acepção dos recursos, através da apresentação do modelo de desempenho do ambiente de provimento de serviços Internet, a definição de um projeto experimental, seus fatores, níveis e métricas de desempenho. Ao final são apresentados os resultados e as respectivas análises.

5.1 Desenvolvimento de um modelo de desempenho

Para o desenvolvimento de um projeto de desempenho é necessário selecionar dentre as três técnicas de avaliação, uma que melhor se adapte ao ambiente em estudo. A técnica utilizada neste trabalho é a de simulação.

A simulação computacional de sistemas consiste na utilização de determinadas técnicas matemáticas empregadas em computadores, as quais permitem imitar o funcionamento de praticamente, qualquer tipo de operação ou processo do mundo real. A simulação busca:

- Descrever o comportamento do sistema;
- Construir teorias e hipóteses considerando as observações efetuadas e,

- Usar o modelo para prover o comportamento futuro, isto é, os efeitos produzidos por alterações no sistema ou nos métodos empregados em sua operação.

5.1.1 Justificativas para o uso de simulação

Dentre os motivos que justificam o uso de simulação pode-se citar:

- Modelos mais realistas: maior liberdade na construção de um modelo. A simulação não obriga a enquadra um problema em determinado molde para que se possa obter uma solução. Assim, no lugar de soluções exatas para problemas aproximados, tem-se soluções aproximadas para problemas mais reais;

- Processo de modelagem evolutivo: permite começar com um modelo simples e aos poucos, ir identificando de maneira mais clara as peculiaridades do problema; em função deste aprendizado, tem-se condições de aperfeiçoar este modelo incorporando novas variáveis;

- Perguntas do tipo "e se": muitas vezes, no lugar de buscar uma solução, o objetivo resume-se em tornar mais claras as possíveis conseqüências de um conjunto de soluções. A simulação é propícia à formulação de perguntas do tipo "e se" ? que permitem avaliar com base no modelo, o efeito de possíveis mudanças de cenário ou de diferentes decisões;

- Facilidade de comunicação: um modelo de comunicação é em geral, muito mais fácil de se compreender do que um conjunto de complicadas equações matemáticas, como é o caso da modelagem analítica.

5.1.2 Vantagens e desvantagens da simulação

Apesar da simulação ser uma excelente ferramenta de análise, é preciso conhecer com mais profundidade as suas vantagens e desvantagens. As vantagens apresentadas são:

- Uma vez criado, um modelo pode ser utilizado inúmeras vezes para avaliar projetos e políticas propostas;
- A metodologia de análise utilizada pela simulação permite a avaliação de um sistema proposto, mesmo que os dados de entrada estejam ainda na forma de esquemas ou rascunhos;
- A simulação é geralmente mais fácil de aplicar do que métodos analíticos;
- Hipóteses sobre como ou por que certos fenômenos acontecem podem ser testadas para confirmação;
- Permite compreender melhor quais variáveis são as mais importantes em relação ao desempenho e como as mesmas interagem entre si e com os outros elementos do sistema;
- A identificação de gargalos é facilitada;

Embora sejam inúmeras as vantagens, o processo de simular apresenta algumas dificuldades, como as listadas a seguir:

- A construção de modelos requer um treinamento especial, principalmente da ferramenta que será utilizada;
- Os resultados da simulação são, muitas vezes, de difícil interpretação. Uma vez que os modelos tentam capturar a aleatoriedade do sistema, muitas vezes existem dificuldades em determinar quando uma observação realizada durante uma execução se deve a alguma significante relação no sistema ou a aleatoriedade construída no modelo;
- A modelagem e a experimentação associadas a modelos de simulação consomem muitos recursos, principalmente tempo.

5.1.3 Modelagem

No encaminhamento de um estudo de simulação a etapa principal consiste na modelagem do sistema sob estudo para que se possa observar seu comportamento sob determinadas condições na busca de sua compreensão. Este processo de imitação e criação de uma história artificial da atuação e desempenho dos sistemas reais pressupõe na maioria das vezes, uma série de simplificações sobre seu funcionamento de forma que se possa, cientificamente, estudá-los e entendê-los.

Existem várias ferramentas disponíveis no mercado que realizam a modelagem do desempenho das aplicações. As mais conhecidas são: BEST/1, PerformanceWorks 3.0, Network II.5, NETClarity, View Point Capacity Planner, COMNET III, ARENA, Resolute Apllication Performance System, OPNET e QASE BACKGRUOND.

5.1.3.1 Ferramenta utilizada para modelagem

Para a realização do modelo de desempenho é utilizada a ferramenta COMNET III, desenvolvida pela empresa CACI Products Company. É uma ferramenta de modelagem específica para redes de computadores que prevê o desempenho com alta fidelidade e precisão de LANs e WANs, permitindo a análise e criação de diferentes alternativas antes de realizar possíveis erros. Utilizada para realizar previsões e instantaneamente mostrar o impacto das mudanças na rede antes de serem implementadas. Ilustra como novas aplicações, novo *hardware* e trocas na largura de banda poderão afetar o desempenho do sistema. Pode-se propor cenários para testes e obter em poucos segundos o resultado em forma de relatórios e gráficos indicando como o desempenho da rede será afetado.

A interface inicial da ferramenta é mostrada na fig. 26.

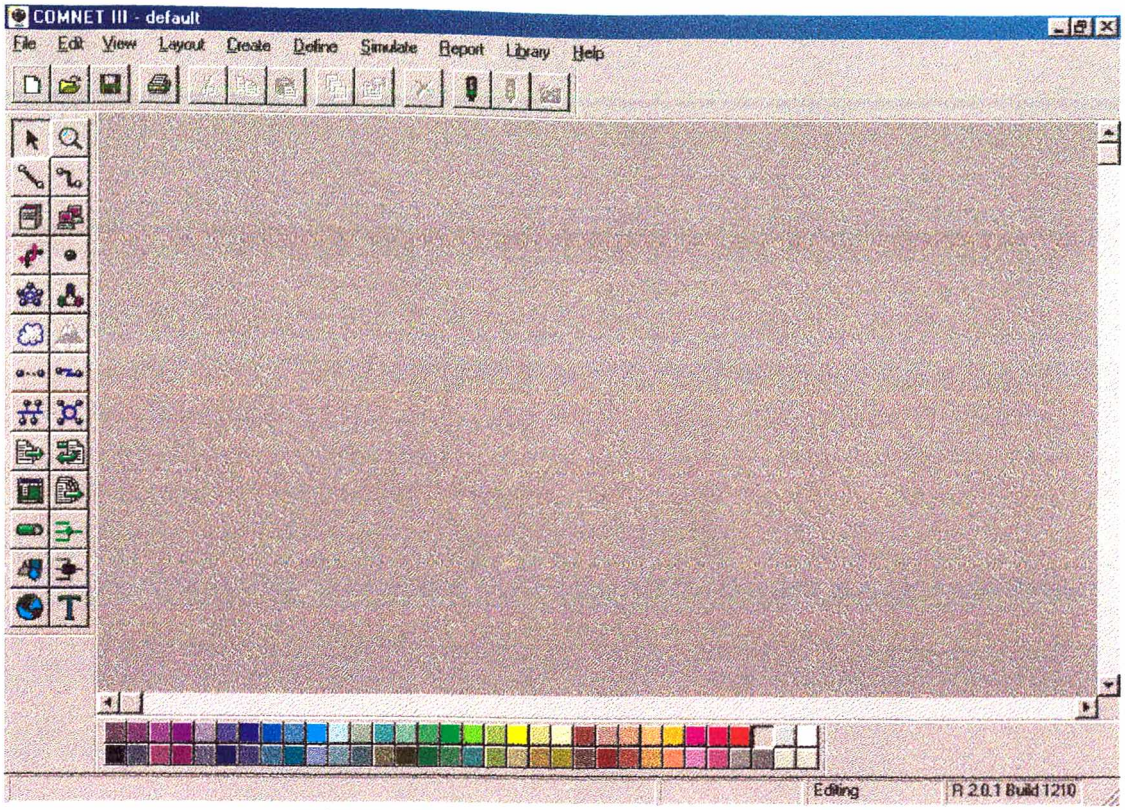


FIGURA 26 – Interface inicial do COMNET III

A ferramenta define a topologia da rede através da utilização de nodos *e links*. Os nodos desempenham as funções de processamento e contém portas que são conectadas aos *links*. Há dois tipos de nodos, os de aplicação e de comunicação (*vide fig. 27 e 28*).

A carga da rede é produzida pelos geradores de mensagens e aplicações (fig. 31 e 33) Um nodo pode ter várias mensagens sendo geradas a partir dele e cada mensagem tem uma distribuição de tamanho, um intervalo entre chegadas e uma destinação. As mensagens geradas são transportadas ao destino através de protolocos de transporte, roteamento e controle de acesso ao meio. O de transporte divide a mensagem em pacotes no nodo e remonta a mensagem no nodo destino, dependendo do protoloco utilizado.

Para ilustração, são apresentados a seguir os componentes antes descritos:

FIGURA 27 – Nodo de aplicação

O nodo de aplicação pode ser usado para modelar sistemas finais, *bridges*, *gateways* e *switches* de comunicação permitindo rotear tráfego através dele.

FIGURA 28 – Grupo de nodos de aplicação

Os grupos de nodos de aplicação são restritos para modelagem de sistemas finais. Somente podem gerar e receber o tráfego não permitindo a passagem através dos mesmos.

FIGURA 29 – Nodo de comunicação

Os nodos de comunicação modelam o *hardware* usado para rotear o tráfego incluindo roteadores, *hubs e switches*. A ferramenta disponibiliza uma biblioteca com inúmeros equipamentos de diversos fabricantes e suas respectivas características que precisam simplesmente ser selecionados. Possibilita também redefinir parâmetros destes equipamentos de acordo como o ambiente em estudo.

FIGURA 30 – Tipos de *Links*

Existem duas classes de *links* disponíveis no COMNET III: ponto-a-ponto, para representar um canal entre dois nodos e o *link* multi-acesso para redes locais e outras simulações onde mais que dois nodos dividem o mesmo meio de comunicação. Os *links* disponíveis no COMNET III modelam vários protocolos, como CSMA/CD, CSMA/CA, *Aloha, Token Ring, Token Bus*, FDDI, entre outros.

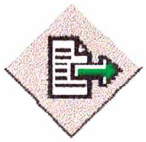


FIGURA 31 – Gerador de Mensagem

FIGURA 32 – Gerador de Resposta

FIGURA 33 – Aplicação

Os três componentes descritos anteriormente são os responsáveis pela carga de trabalho que circula no sistema.

Um modelo é construído através da execução dos seguintes passos:

- Nodos, *links* e recursos de carga são selecionados da barra de ferramentas;
- Estes elementos são conectados e definidos os seus relacionamentos;
- Os parâmetros de cada elementos são ajustados de acordo com o ambiente;
- Operações da rede e parâmetros de protocolos são setados;
- O modelo é verificado e executado e os resultados são apresentados em relatórios.

5.1.3.2 Modelagem do Ambiente

Na realização da modelagem foi levado em consideração apenas o ambiente do provedor pois não tem como exercer influência sobre o comportamento e desempenho da rede Internet. As respostas a partir do *link* com a empresa Brasil Telecom serão vistas como uma espécie de "caixa preta". A figura 34 representa este fato, ou seja, o provedor de serviços somente exerce influência para a melhora no desempenho ao usuário final nos componentes de seu ambiente.

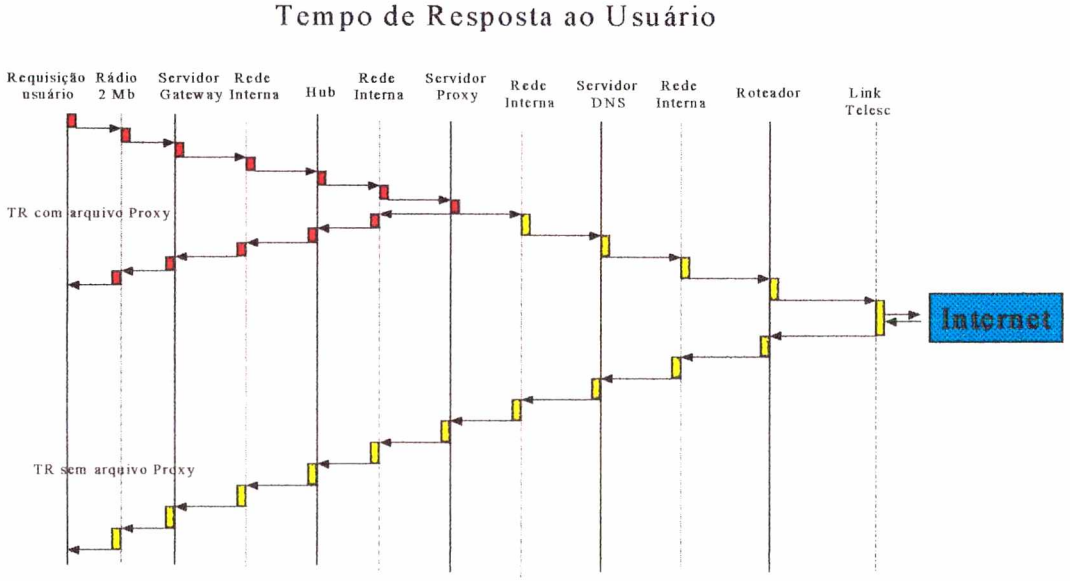


FIGURA 34 – Tempo de resposta ao usuário

O ambiente foi modelado com base na topologia e a descrição do ambiente relacionados no capítulo 4. Algumas simplificações foram realizadas:

- Clientes de acesso discado não são considerados porque a ferramenta utilizada é uma versão acadêmica e somente permite utilizar no máximo 20 componentes;

- Dados como velocidade do processador e memória dos servidores não podem ser incluídos no nodo de aplicação quando se utiliza o módulo gerador de mensagens (é o caso do modelo realizado). Como os pacotes que trafegam pelos servidores geram processamento nestes equipamentos foi adicionado um *delay* em cada servidor para representar este processo;

- O servidor *proxy* não é modelado como ele realmente funciona, ou seja, ler e armazenar arquivos localmente. Ele é modelado apenas como um nodo de aplicação com um *delay* associado.

Para a determinação do período de simulação foram adotados os seguintes procedimentos:

- Realização de duas rodadas de simulação. A primeira correspondente a 600 segundos e a segunda a 1800 segundos. Foram observadas poucas diferenças em termos de respostas do sistema. Os resultados apresentados a seguir referem-se a segunda simulação;

- Observação visual dos gráficos que mostram o comportamento das variáveis de desempenho objetivando determinar o período transiente e de estabilidade. Um dos gráficos pode ser visualizado na fig 35.

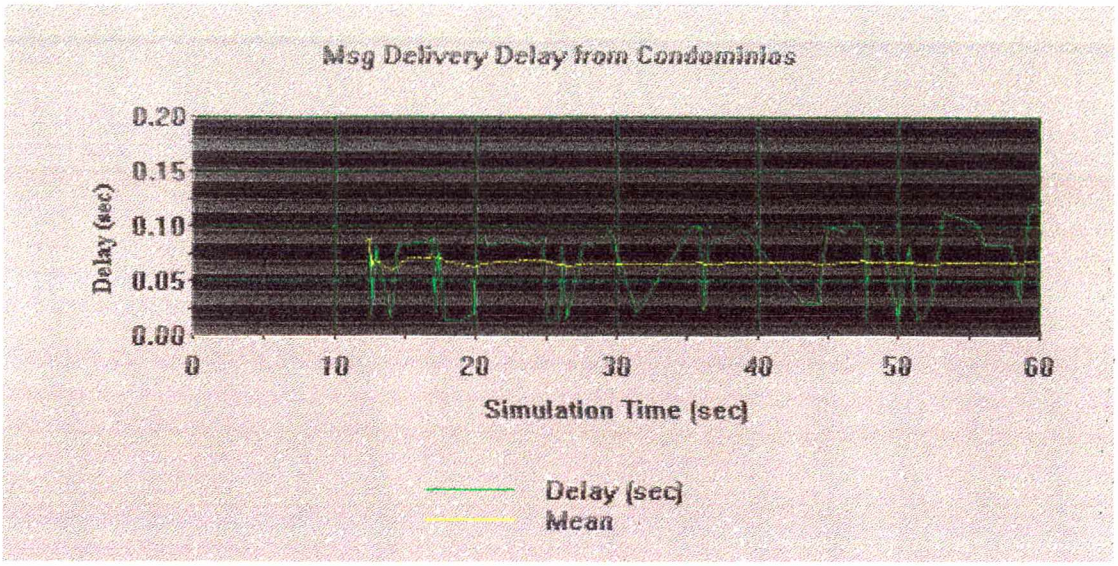


FIGURA 35 – Período de *warm-up*

Observou-se que a estabilidade ocorre após 20 segundos de simulação, período adotado como *warm-up*[1]. Desta forma, os dados utilizados na análise dos resultados, foram apenas após este período.

O modelo de desempenho é apresentado na fig 36 e representa as requisições dos usuários dos condomínios e da Empresa 1 com as respectivas respostas a partir dos servidores Internet e servidor *proxy*, conforme o caso. As cargas foram descritas no capítulo anterior e foram adicionadas ao modelo. Os clientes de acesso discado não puderam ser incluídos no modelo em função da limitação da quantidade de componentes da ferramenta.

Os resultados da simulação foram validados com os dados observados no ambiente real comparando-se o percentual de utilização de um dos recursos: o *link* de 512 Kb. Como pode ser observado na fig. 19, capítulo 4, no horário de pico o *link* está

---

[1] *Warm-up*: é conhecido como aquele tempo inicial onde o sistema não está normalizado. Não possuem condições iniciais fixas até que o sisterma entre em normalidade.

com uma taxa de ocupação em torno de 90%. O modelo de desempenho apresenta uma ocupação de 91,80% com o modelo de carga definido.

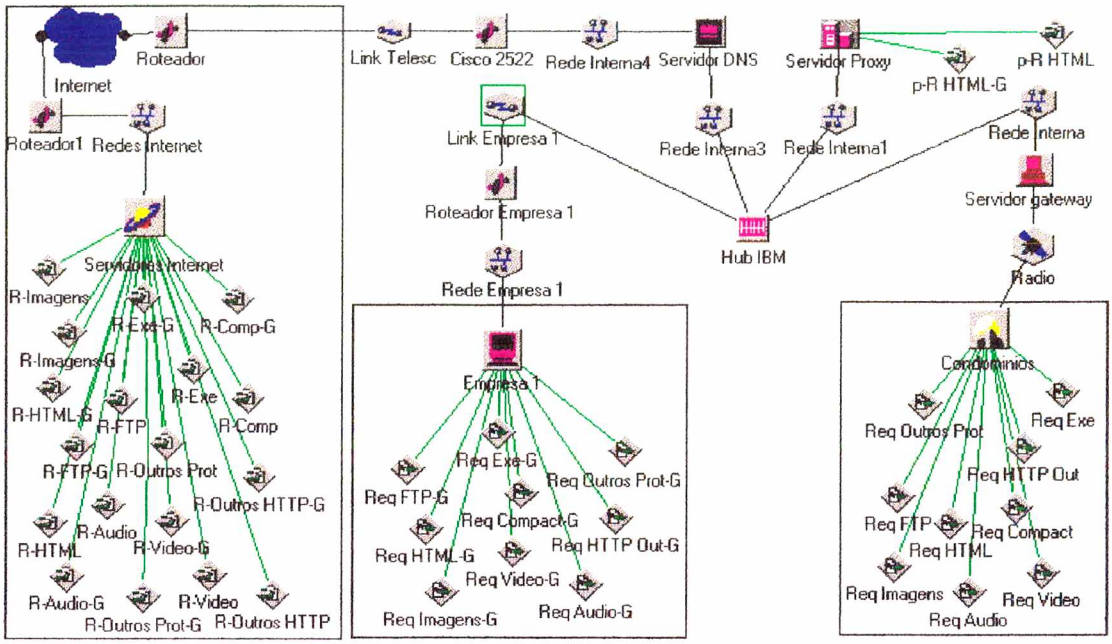


FIGURA 36 – Modelo de desempenho

5.2 Projeto Experimental

Para a realização do projeto experimental é necessário inicialmente identificar quais os fatores, métricas e tipo de projeto a ser utilizado. A descrição de cada uma destas etapas é feita a seguir.

5.2.1 Seleção dos fatores para desempenho

Antes da escolha dos fatores é importante fazer uma lista completa das características e cargas de trabalho que afetam o desempenho do sistema. Estas características são chamadas de parâmetros. Em seguida, escolhe-se dentre esta lista, os mais importantes, ou seja, aqueles que se variarem, terão impactos significantes sobre o

desempenho. Para o ambiente em estudo pode-se definir os seguintes parâmetros que afetam o desempenho:

- Velocidade da CPU dos servidores;
- Memória dos servidores;
- Tamanho do HD dos servidores;
- Velocidade da rede de comunicação;
- Overhead do sistema na interface com o canal;
- Overhead do sistema na interface da rede;
- Velocidade do roteadores;
- Largura de banda do *link* externo;
- Política de substituição de arquivos do *proxy*;
- Esquema de armazenamento de arquivos no *proxy*;
- Quantidade de clientes;
- Capacidade do Hub.

Para o ambiente em estudo foram identificados e serão utilizados os fatores e níveis descritos a seguir.

5.2.1.1 Fator 1 – Largura de banda da Rede

Este fator foi o primeiro definido pela importância do mesmo no ambiente, pois todas as requisições não atendidas no servidor *proxy* utilizam este recurso. Além disso, através da análise dos gráficos de utilização do mesmo, observa-se que a capacidade atual já está sendo quase que totalmente utilizada. Os níveis definidos são:

- Nível (-): 512 Kb
- Nível (+): 1024 Kb

5.2.1.2 Fator 2 - Quantidade de clientes condôminos

Este fator implica diretamente na variação da intensidade da carga de trabalho e determina a quantidade de requisições em trânsito no sistema. Aumentando a

quantidade deste tipo de clientes de acordo com o plano de negócios da organização espera-se prever o comportamento do ambiente processando a carga futura. Desta forma é possível avaliar quais componentes suportarão esta nova carga e quais apresentarão gargalos, impactando na qualidade de serviços oferecida aos clientes.

Para efeito de simulação considera-se que no horário de pico em torno de 20% dos clientes estarão utilizando os serviços ao mesmo tempo.

Quantidade de clientes:

- Nível (-): 150 clientes * 20%: 30 clientes
- Nível (+): 240 clientes * 20%: 48 clientes

De acordo com o planejamento estratégico da empresa aumenta-se em torno de 15 novos clientes condôminos a cada mês. Fazendo-se uma projeção para um horizonte de seis meses pode-se analisar o ambiente processando a carga que será gerada naquele período. Os 240 clientes representam este horizonte.

5.2.1.3 Fator 3 – Taxa de acerto em função da política de substituição de arquivos do servidor *Proxy*

O servidor *proxy* é utilizado por todas as requisições que transitam pelo sistema e que requerem um processamento de pesquisa para saber se o arquivo solicitado está ou não armazenado no local.

Conforme descrito no item 3.3.7, a política de substituição de arquivos tem grande importância na otimização do ambiente porque dependendo da técnica utilizada, permite aumentar o número de requisições atendidas pelo servidor, diminuir o volume de dados que trafegam na rede otimizando a utilização da largura de banda e reduzir o tempo de resposta para o usuário final.

Pensou-se em utilizar estas políticas como fatores no ambiente, entretanto, por limitação da ferramenta de modelagem, isto não foi possível. Em função disto, foi

utilizada uma simplificação considerando que uma política que não otimize estes aspectos tenha um percentual de acerto menor do que as demais.

A empresa utiliza a política LFU e conforme a tabela 4, há outras políticas mais eficientes. Portanto, serão utilizados os seguintes níveis, considerando-se que 20% representa a política de substituição de arquivos atual e 40% uma política mais otimizada:

- Nível (-): 20%
- Nível (+): 40%

5.2.1.4 Fator 4 – Política de armazenamento de arquivos no servidor *Proxy*

A empresa utiliza como política manter na *cache* do servidor *proxy* somente os arquivos HTML requisitados pelos clientes. Entretanto conforme descrito na definição da carga de trabalho verifica-se que a maioria dos arquivos solicitados são imagens e outros arquivos do protocolo HTTP. Utiliza-se então este fato como um dos fatores para verificar o impacto da mudança e adoção de outro esquema de armazenamento dos arquivos, de maneira a otimizar o tempo de resposta:

Política de armazenamento do servidor proxy:
- Nível (-): arquivos HTML;
- Nível (+): arquivos HTML, Imagens, Áudio e Outros HTTP (conforme carga definida);

Não se sugere o armazenamento dos arquivos de vídeo, compactado FTP e Outros Protocolos em função dos seus tamanhos e baixa freqüência de requisição.

5.2.2 Seleção das métricas

A avaliação de desempenho é feita através de métricas (também chamadas de variáveis de interesse ou de desempenho), que são critérios para a comparação de

desempenho entre sistemas, ou entre diferentes situações a que um sistema pode ser submetido. Para qualquer estudo, um conjunto de métricas de desempenho deve ser escolhido.

Segundo MENASCÉ & ALMEIDA (1998), a percepção de desempenho na web pode ser analisada de duas formas, dependendo do ponto de vista. Para o usuário o importante é o tempo de resposta e conexões não recusadas. Para o administrador é o alto *throughput* e a disponibilidade dos recursos.

Segundo RENAUD (1994), o tempo de resposta é uma medida de quantidade de tempo necessária para realizar o trabalho. É estimado ou medido por um certo número de transações e expresso em termos de tempo médio por transação. A definição de uma transação depende da aplicação, mas o tempo de resposta é sempre medido desde a hora em que o usuário inicia um pedido até a hora em que os resultados são apresentados ao usuário. Tendo em vista a sua importância para o usuário final, o objetivo maior é minimizá-lo.

O tempo de resposta para os usuários no ambiente web inclui latência no servidor mais o tempo despendido na comunicação na rede, mais o tempo de processamento na máquina do cliente. Portanto, o desempenho percebido pelo usuário depende da capacidade do servidor, da carga da rede, da largura de banda e da máquina do cliente.

No caso do administrador do sistema, as métricas mais importantes no caso do ambiente web são a latência, *throughput* e disponibilidade dos recursos. A taxa de quantas requisições são servidas representa o *throughput*. Devido a larga variabilidade de tamanho dos objetos web requisitados, o *throughput* geralmente é medido em termos de megabits por segundo (Mbps). O tempo requerido para completar uma requisição é a latência, o qual é um dos componentes do tempo de resposta ao cliente. A latência média no servidor é a média de tempo para executar as requisições. A disponibilidade dos recursos refere-se a taxa de ocupação dos mesmos.

È importante lembrar que as métricas são as peças chave de todo o processo de avaliação de desempenho. Analisando o ambiente em estudo e considerando as observações anteriores, as métricas utilizadas no trabalho são:

- Tempo de resposta ao usuário final: média de todos os tipos de arquivos;
- Percentual de utilização do *link* externo.

5.2.3 Experimentação

O objetivo de um projeto experimental formal é obter o máximo de informações com um mínimo de experimentos. Basicamente, num projeto experimental deseja-se:

- Projetar um conjunto de experimentos para a simulação;
- Desenvolver um modelo que melhor descreva os dados obtidos;
- Estimar a contribuição de cada alternativa;
- Estimar o intervalo de confiança de cada alternativa;
- Verificar se as alternativas apresentam diferenças significantes e se o modelo é adequando.

Os seguintes termos são freqüentemente usados na análise e projetos de experimentos:

- **variável de resposta**: é o resultado de um experimento. Geralmente a variável de resposta é a medida de desempenho do sistema;

- **fatores**: cada variável que afeta a variável de resposta e que pode assumir diferentes valores relevantes;

- **fatores primários**: os fatores cujos efeitos necessitam ser quantificados;

- **fatores secundários**: fatores que incidem no desempenho mas cuja incidência não estamos interessados em quantificar.

- **níveis**: são os valores que os fatores podem assumir. Em outras palavras, cada nível do fator constitui uma alternativa para aquele fator;

- **replicação**: repetição de todos ou alguns experimentos;

- **projeto experimental:** consiste da especificação do número de experimentos, as combinações do nível do fator para cada experimento e o número de replicações de cada experimento.

5.2.3.1 Tipos de projeto experimental

De uma forma geral a decisão sobre as combinações a serem usadas deve ser tomada de acordo como o número de fatores, quantidade de níveis de cada um, tempo e recursos disponíveis para os testes. Os projetos mais freqüentemente usados são projeto simples, projeto fatorial completo, projeto fatorial fracionário e projeto fatorial 2^k. Segue a seguir, uma explicação de cada um destes projetos.

5.2.3.1.1 Projeto simples

Esta abordagem começa com uma configuração típica, variando um fator por vez para ver como este fator afeta no desempenho. Este projeto somente pode ser utilizado se não houver interações entre os fatores, pois pode levar a conclusões erradas. Para K fatores, com o ith fator tendo n_i níveis, requer n experimentos, onde:

$$n = 1 + \sum_{i=1}^{k} (n_i - 1)$$

5.2.3.1.2 Projeto Fatorial Completo

Utiliza todas as combinações possíveis de todos os níveis de cada fator. É o que gera mais informações, mas só é factível se o número de fatores e de níveis em cada um não for alto e quando não for necessária a realização de várias replicações de um mesmo experimento.

A vantagem de um projeto fatorial completo é que todas as combinações possíveis de configuração e carga de trabalho serão examinadas. Pode-se encontrar o efeito de todos os fatores, incluindo os fatores secundários e suas interações. O principal problema é o custo do estudo, pois pode tomar muito tempo e dinheiro para conduzir todos estes experimentos, que são repetidos várias vezes. Um desvantagem é o rápido crescimento da quantidade de experimentos necessários quando temos um crescimento do número de fatores ou fatores com muitos níveis. Um caso especial é quando todos os fatores possuem apenas dois níveis. Assim, com k fatores, teremos 2^k experimentos. Neste caso, o projeto é chamado de projeto fatorial completo 2^k.

Existem algumas alternativas para reduzir o número de experimentos:

- reduzir o número de níveis por fator;
- reduzir o número de fatores;
- usar projetos fatoriais fracionários;

Para K fatores, com o ith fator tendo n_i níveis, requer n experimentos, onde:

$$n = \prod_{i=1}^{k} n_i$$

5.2.3.1.3 Projeto Fatorial Fracionário

Às vezes o número de experimentos requeridos para um projeto fatorial completo é muito grande, sendo que isto ocorre quando se tem uma grande quantidade de fatores ou níveis. Esta abordagem usa apenas uma parte das combinações que seriam

usadas num fatorial completo, reduzindo o número de fatores e o número de níveis. segundo certos critérios. Exige menos testes, mas em contrapartida também gera menos informações. Entretanto, para cada vantagem existe uma desvantagem. É possível que não se obtenha todas as interações e, portanto, seus efeitos sobre a resposta.

5.2.3.1.4 Projeto Fatorial 2^k

Um projeto experimental com 2^k é usado para determinar o efeito de k fatores, cada um dos quais com duas alternativas ou níveis. O primeiro passo é escolher aqueles fatores que tem um impacto significante sobre o desempenho do sistema.

Para a análise de desempenho do ambiente em estudo foi escolhido este projeto, principalmente em função das características:

- Tem-se quatro fatores importantes que ser quer avaliar;
- Cada fator tem dois níveis;
- Permite a ordenação dos fatores na ordem do impacto.

5.3 Análise dos resultados da simulação

A partir do modelo de desempenho descrito acima foram realizadas 16 simulações variando-se os fatores e níveis conforme descrito na tabela 21.

Fatores	Níveis	
	(-)	(+)
Capacidade do link	512 Kb	1024 Kb
Quantidade de clientes	30	48
Política substituição arquivos *Proxy*	20%	40%
Política armazenamento *Proxy*	Somente HTML	outros

TABELA 21 – Fatores utilizados na simulação

Os resultados obtidos nas simulações em termos de métricas são representados na tabela 22 com as seguintes siglas:

- UL: percentual de utilização do *link* externo;
- TR: tempo médio de resposta em segundos ao usuário final;

Experimentos	Link	Qtde	% Acerto	Política	UL	TR
1	1024	48	40	Outros	77,96	131
2	1024	48	40	HTML	74,93	80
3	1024	48	20	Outros	62,97	50
4	1024	48	20	HTML	65,97	66
5	1024	30	40	Outros	30,77	60
6	1024	30	40	HTML	44,08	52
7	1024	30	20	Outros	34,30	38
8	1024	30	20	HTML	49,00	55
9	512	48	40	Outros	90,57	222
10	512	48	40	HTML	100,00	741
11	512	48	20	Outros	100,00	682
12	512	48	20	HTML	100,00	798
13	512	30	40	Outros	59,34	47
14	512	30	40	HTML	93,08	130
15	512	30	20	Outros	75,25	71
16	512	30	20	HTML	91,80	166

TABELA 22 – Resultados da simulação

A partir da tabela 22 várias análises podem ser feitas. Pode-se analisar os resultados das simulações aos pares, ou através da distribuição da variação para determinar quais são os fatores que exercem mais influência no desempenho do sistema. Os dois tipos de análises estão descritos nas seções seguintes.

5.3.1 Distribuição da variação

Abaixo descreve-se os efeitos de cada fator calculados por métrica definida anteriormente, ou seja, tempo de resposta e percentual de utilização do *link*.

Para as duas análises os seguintes dados são considerados:

Fator	Tipo
A	Capacidade do *link*
B	Quantidade de clientes
C	Política de substituição - *proxy*
D	Política de armazenamento - *proxy*

TABELA 23 – Fatores

-1	1
512	1024
30	48
20%	40%
HTML	Outros

TABELA 24 – Níveis

5.3.1.1 Métrica - Tempo de Resposta

Para o cálculo dos efeitos de cada fator faz-se necessário definir a tabela de sinais, conforme a seguir. A importância de um fator é medida pela proporção do total da variação que este representa.

EXP	I	A	B	C	D	AB	AC	AD	BC	BD	CD	TR
1	1	1	1	1	1	1	1	1	1	1	1	131,41
2	1	1	1	1	-1	1	1	-1	1	-1	-1	79,76
3	1	1	1	-1	1	1	-1	1	-1	1	-1	49,77
4	1	1	1	-1	-1	1	-1	-1	-1	-1	1	65,75
5	1	1	-1	1	1	-1	1	1	-1	-1	1	59,77
6	1	1	-1	1	-1	-1	1	-1	-1	1	-1	52,18
7	1	1	-1	-1	1	-1	-1	1	1	-1	-1	38,39
8	1	1	-1	-1	-1	-1	-1	-1	1	1	1	54,99
9	1	-1	1	1	1	-1	-1	-1	1	1	1	221,92
10	1	-1	1	1	-1	-1	-1	1	1	-1	-1	741,89
11	1	-1	1	-1	1	-1	1	-1	-1	1	-1	682,28
12	1	-1	1	-1	-1	-1	1	1	-1	-1	1	798,35
13	1	-1	-1	1	1	1	-1	-1	-1	-1	1	46,91
14	1	-1	-1	1	-1	1	-1	1	-1	1	-1	129,65
15	1	-1	-1	-1	1	1	1	-1	1	-1	-1	71,19
16	1	-1	-1	-1	-1	1	1	1	1	1	1	165,99
	3.390,20	-2.326,16	2.152,06	-463,22	-786,92	-1.909,34	691,66	840,24	-379,12	-413,82	-300,02	
	211,89	-145,39	134,50	-28,95	-49,18	-119,33	43,23	52,52	-23,70	-25,86	-18,75	

TABELA 25 – Tabela dos Sinais – Tempo de Resposta

Em seguida a equação denominada Soma do Total dos Quadrados é aplicada com o objetivo de identificar a alocação da variação de cada fator:

SQT = SQA + SQB + SQC + SQD + SQAB + SQAC + SQAD + SQBC + SQBD + SQCD

Na análise da métrica tempo de resposta obteve-se a seguinte distribuição de variação:

SQA =	1.014.566,31	Capacidade do link (A)	33,59%
SQB =	868.380,42	Quantidade de clientes (B)	28,75%
SQC =	40.232,39	Política substituição proxy (D)	1,33%
SQD =	116.108,08	Política armazenamento proxy (D)	3,84%
SQAB =	683.546,11	Interação entre A e B	22,63%
SQAC =	89.698,79	Interação entre A e C	2,97%
SQAD =	132.375,61	Interação entre A e D	4,38%
SQBC =	26.949,75	Interação entre B e C	0,89%
SQBD =	32.108,81	Interação entre B e D	1,06%
SQCD =	16.877,25	Interação entre C e D	0,56%
SQT =	3.020.843,52		100,00%

TABELA 26 – Distribuição de variação – Tempo de Resposta

O percentual de variação atribuído a cada um dos fatores relacionados na tabela 26 ajuda a decidir o quanto um determinado fator é importante na medida do impacto que causa sobre a métrica. Pode-se concluir que:

- A capacidade do link é o fator que possui mais influência no tempo de resposta, sendo responsável por 33,59% da variação;
- A intensidade da carga é responsável por 28,75% da variação;
- Existe uma grande interação entre os dois fatores acima, que representa 22,63% do total da variação, quer dizer, a variação de um influencia diretamente no resultado do segundo;
- O tempo de resposta médio obtido para todas as simulações é de 211 segundos.

5.3.1.2 Métrica – Utilização do Link

A tabela de sinais para a métrica utilização do *link* é:

EXP	I	A	B	C	D	AB	AC	AD	BC	BD	CD	% Link
1	1	1	1	1	1	1	1	1	1	1	1	77,96
2	1	1	1	1	-1	1	1	-1	1	-1	-1	74,93
3	1	1	1	-1	1	1	-1	1	-1	1	-1	62,97
4	1	1	1	-1	-1	1	-1	-1	-1	-1	1	65,97
5	1	1	-1	1	1	-1	1	1	-1	-1	1	30,77
6	1	1	-1	1	-1	-1	1	-1	-1	1	-1	44,08
7	1	1	-1	-1	1	-1	-1	1	1	-1	-1	34,30
8	1	1	-1	-1	-1	-1	-1	-1	1	1	1	49,00
9	1	-1	1	1	1	-1	-1	-1	1	1	1	90,57
10	1	-1	1	1	-1	-1	-1	1	1	-1	-1	100,00
11	1	-1	1	-1	1	-1	1	-1	-1	1	-1	100,00
12	1	-1	1	-1	-1	-1	1	1	-1	-1	1	100,00
13	1	-1	-1	1	1	1	-1	-1	-1	-1	1	59,34
14	1	-1	-1	1	-1	1	-1	1	-1	1	-1	93,08
15	1	-1	-1	-1	1	1	1	-1	1	-1	-1	75,25
16	1	-1	-1	-1	-1	1	1	1	1	1	1	91,80
	1.150,02	-270,06	194,78	-8,56	-87,70	52,58	39,56	31,74	37,60	68,90	-19,20	
	71,88	-16,88	12,17	-0,53	-5,48	3,29	2,47	1,98	2,35	4,31	-1,20	

TABELA 27 – Tabela dos Sinais – Utilização do link

Na análise da métrica Utilização do *Link* obteve-se a seguinte distribuição de variação:

SQA =	13674,8	Capacidade do link (A)	55,89%
SQB =	7113,61	Quantidade de clientes (B)	29,07%
SQC =	13,7388	Política substituição proxy (C)	0,06%
SQD =	1442,12	Política armazenamento proxy (D)	5,89%
SQAB =	518,373	Interação entre A e B	2,12%
SQAC =	293,436	Interação entre A e C	1,20%
SQAD =	188,893	Interação entre A e D	0,77%
SQBC =	265,08	Interação entre B e C	1,08%
SQBD =	890,102	Interação entre B e D	3,64%
SQCD =	69,12	Interação entre C e D	0,28%
SQT =	24469,3		100,00%

TABELA 28 – Distribuição de variação – Utilização do Link

A partir da tabela anterior pode-se concluir:

- O fator que possui mais influência na utilização do link também é a capacidade do canal, sendo responsável por 55,89% da variação;
- A quantidade de clientes é responsável por 29,07% da variação;
- A utilização média do link em todas as simulações é de 71,88%;
- A variação na política de armazenamento no proxy exerce mais influência na utilização do link que a política de substituição de arquivos.

5.3.2 Análise de cenários possíveis

Os resultados apresentados anteriormente através da alocação da variação parecem um tanto óbvios, ou seja, é claro que a capacidade do canal exerce a maior influência no tempo de resposta e no percentual de utilização do canal. Busca-se então nesta seção, analisar os resultados das simulações aos pares para identificar a influência dos demais fatores e sugerir cenários alternativos.

5.3.2.1 Cenário 1

Este cenário considera no nível 1 o ambiente atual processando a carga atual. No nível 2 verifica os resultados em termos das métricas definidas, caso fosse feita uma alteração na política de armazenamento de arquivos.

	Fatores				Resultados	
Níveis	Link	Qtde	% Proxy	Política	% Link	Tempo Resposta
1	512	30	20%	HTML	91,80	166
2	512	30	20%	Outros	75,25	71

TABELA 29 – Cenário 1

Através da mudança na política de armazenamento de arquivos no servidor *proxy*, os resultados da simulação mostram que o desempenho já melhoraria significativamente em termos das duas métricas. O mais importante a se observar é que para se obter estes resultados não há a necessidade de dispêndio de recursos financeiros.

5.3.2.2 Cenário 2

Este cenário considera o ambiente atual processando a carga atual e sugere uma alteração na política de substituição de arquivos.

	Fatores				Resultados	
Níveis	Link	Qtde	% Proxy	Política	% Link	Tempo Resposta
1	512	30	20%	HTML	91,80	166
2	512	30	40%	HTML	93,08	129

TABELA 30 – Cenário 2

Os resultados mostram que não há ganhos significativos em termos de % de utilização do *link*, entretanto o tempo de resposta apresenta um ganho razoável, que poderia justificar a mudança da configuração atual.

5.3.2.3 Cenário 3

Este cenário considera o ambiente atual processando a carga atual e sugere uma alteração nas políticas de substituição e armazenamento dos arquivos no servidor *proxy*.

	Fatores				Resultados	
Níveis	Link	Qtde	% Proxy	Política	% Link	Tempo Resposta
1	512	30	20%	HTML	91,80	166
2	512	30	40%	Outros	59,34	47

TABELA 31 – Cenário 3

Melhor dos cenários apresentados até o momento pois reduz significativamente as duas métricas e não necessita injeção de recursos financeiros no ambiente, bastando trocas nas configurações atuais.

5.3.2.4 Cenário 4

Este cenário considera o ambiente atual processando uma previsão de carga futura.

	Fatores				Resultados	
Níveis	Link	Qtde	% Proxy	Política	% Link	Tempo Resposta
1	512	30	20%	HTML	91,80	166
2	512	48	20%	HTML	100,00	798

TABELA 32 – Cenário 4

Através da análise da tabela pode-se concluir que se mantidas as condições atuais a qualidade dos serviços prestados pelo provedor decairão sensivelmente num horizonte de seis meses.

5.3.2.5 Cenário 5

Este cenário considera o ambiente atual processando uma previsão de carga futura e sugere uma alteração na política de armazenamento.

	Fatores				Resultados	
Níveis	Link	Qtde	% Proxy	Política	% Link	Tempo Resposta
1	512	30	20%	HTML	91,80	166
2	512	48	20%	Outros	100,00	682

TABELA 33 – Cenário 5

Os resultados mostram que não se tem ganhos significativos nestas condições.

5.3.2.6 Cenário 6

Este cenário considera o ambiente atual processando uma previsão de carga futura e sugere uma alteração na política de substituição de arquivos.

	Fatores				Resultados	
Níveis	Link	Qtde	% Proxy	Política	% Link	Tempo Resposta
1	512	30	20%	HTML	91,80	166
2	512	48	40%	HTML	100,00	741

TABELA 34 – Cenário 6

Também não se apresentam ganhos significativos nestas condições.

5.3.2.7 Cenário 7

Este cenário considera o ambiente atual processando uma previsão de carga futura e sugere uma alteração na política de substituição e armazenamento de arquivos.

	Fatores				Resultados	
Níveis	Link	Qtde	% Proxy	Política	% Link	Tempo Resposta
1	512	30	20%	HTML	91,80	166
2	512	48	40%	Outros	90,57	222

TABELA 35 – Cenário 7

Este resultado é interessante porque mostra que o ambiente atual suporta um aumento de carga se foram feitas alterações nos parâmetros internos. Não é necessário investir na compra de mais capacidade de canal num período próximo, mantida a previsão de crescimento do número de clientes.

5.3.2.8 Cenário 8

O cenário compara o ambiente atual processando a carga atual com um cenário e carga semelhante, apenas alterando a capacidade do canal.

	Fatores				Resultados	
Níveis	Link	Qtde	% Proxy	Política	% Link	Tempo Resposta
1	512	30	20%	HTML	91,80	166
2	1024	30	20%	HTML	49,00	55

TABELA 36 – Cenário 8

O aumento da capacidade do canal, conforme descrito 5.3.1, é um dos fatores que mais exerce influencia nas duas métricas.

5.3.2.9 Cenários 9, 10 e 11

Os cenários abaixo fazem uma comparação da carga atual com a variação dos fatores *link*, política de armazenamento e substituição de arquivos.

	Fatores				Resultados	
Níveis	Link	Qtde	% Proxy	Política	% Link	Tempo Resposta
1	512	30	20%	HTML	91,80	166
2	1024	30	20%	Outros	34,30	38

TABELA 37 – Cenário 9

	Fatores				Resultados	
Níveis	Link	Qtde	% Proxy	Política	% Link	Tempo Resposta
1	512	30	20%	HTML	91,80	166
2	1024	30	40%	HTML	44,08	52

TABELA 38 – Cenário 10

	Fatores				Resultados	
Níveis	Link	Qtde	% Proxy	Política	% Link	Tempo Resposta
1	512	30	20%	HTML	91,80	166
2	1024	30	40%	Outros	30,77	59

TABELA 39 – Cenário 11

Os resultados mostram como a capacidade do canal exerce influência nas duas métricas.

5.3.2.10 Cenário 12

Este cenário faz uma comparação considerando-se a carga futura e variando-se apenas o *link*. Mostra novamente a importância deste fator nas métricas.

	Fatores				Resultados	
Níveis	Link	Qtde	% Proxy	Política	% Link	Tempo Resposta
1	512	48	20%	HTML	100,00	798
2	1024	48	20%	HTML	65,97	66

TABELA 40 – Cenário 12

5.3.2.11 Análise dos cenários

Analisando os vários cenários chega-se a conclusão a melhor alternativa para atender a carga atual e futura é aumentar o link para 1024 Kb. Entretanto esta solução envolve custos muito altos para a sua implantação.

Os cenários também mostram que mantendo a capacidade atual ganhos significativos podem ser obtidos pela simples mudança na política de armazenamento e substituição dos arquivos do servidor *proxy* (cenário 3). E o mais importante é que estas mudanças não geram necessidade de dispêndio de recursos financeiros, apenas mudança nas configurações do ambiente atual. Mesmo aumentando a carga de trabalho para um horizonte de seis meses, este ambiente manterá as métricas em níveis aceitáveis, semelhantes aos atuais, como pode ser visto no cenário 7.

Os cenários comprovam ainda que quando o *link* é aumentado para 1024 Kb a variação dos demais fatores não exerce tanta influência sobre as métricas.

5.4 Previsão do outros ambientes

A previsão é a chave para o planejamento de capacidade porque é necessário ser capaz de determinar como o sistema irá reagir quando ocorrerem trocas na carga e no comportamento dos consumidores ou quando novos modelos de negócios são desenvolvidos. Para prever novos níveis de serviço e diferentes comportamentos dos usuários é necessário fazer suposições a partir do conhecimento do mercado, perspectivas futuras, utilizando o modelo de desempenho.

Levando estes fatos em conta, foram realizadas mais algumas simulações com o modelo de desempenho considerando outros cenários possíveis de ocorrer, conforme descrito nas seções seguintes.

5.4.1 Mudança no comportamento dos consumidores

Como citado por ARLITT *et al* (1999), a mudança na carga de trabalho não ocorre somente pelo aumento de novos usuários mas também devido ao crescimento de uso dos que já são clientes que passam a se acostumar com os serviços e passam a utilizá-lo com maior freqüência. Em função disso, foi simulado um ambiente as seguintes características:

- Link: 512 Kb
- Percentual utilização proxy: 40%
- Política de armazenamento: outros arquivos
- Quantidade de clientes: 150 clientes (quantidade atual), sendo que 40% destes estão utilizando os serviços no mesmo horário, ou seja, 60 usuários (e não apenas 20% como foi modelado a carga atual).

Os resultados da simulação deste ambiente mostram que a utilização do canal ficaria em 100% e o tempo de resposta em 567 segundos. Portanto, se alterar o comportamento dos usuários, mesmo fazendo uma mudança nas políticas de armazenamento e substituição de arquivos atual, os resultados ao usuário final ficariam comprometidos na sua qualidade. Uma comparação dos resultados pode ser visto na tabela a seguir.

	Fatores				Resultados	
Níveis	Link	Qtde	% Proxy	Política	% Link	Tempo Resposta
1	512	30	40%	Outros	59,34	47
2	512	60	40%	Outros	100,00	567

TABELA 41 – Cenário 13

5.4.2 Disponibilização de um novo serviço

Pode-se supor que uma nova tecnologia é criada, o que é muito comum neste ambiente, e que os usuários comecem a requisitá-la. Simula-se então esta situação inserindo uma nova requisição com um tempo entre chegadas Expo(30) e com um

tamanho médio de resposta de 10.000 bytes. O ambiente utilizado possui as seguintes características:

- Link: 512 Kb;
- Percentual utilização proxy: 40%;
- Política de armazenamento: outros arquivos mais o serviço novo;
- Quantidade de clientes: 30 clientes.

	Fatores				Resultados	
Níveis	Link	Qtde	% Proxy	Política	%-Link	Tempo Resposta
1	512	30	40%	Outros	59,34	47
2	512	30	40%	Outros + Novo	100,00	1.012

TABELA 42 – Cenário 14

Através da análise da tabela anterior verifica-se que este novo serviço implicaria em tempos de resposta muito maior. Para atender esta nova demanda alterações no ambiente devem ser realizadas.

5.5 Considerações finais

Ter o último e mais rápido equipamento nem sempre garante o melhor desempenho e isto em muitos casos também não é economicamente viável. Como alternativa as organizações podem obter incremento de desempenho selecionando a correta combinação de equipamentos e cuidadosa otimização dos recursos disponíveis.

Este capítulo mostrou como o planejamento de capacidade pode ajudar neste processo com a realização de simulações e determinação de cenários. Através de avaliações com bases técnicas estes cenários podem ser analisados para determinar quais os mais possíveis de ocorrerem, além possibilitar a sugestão de mudanças em parâmetros procurando-se a melhor relação de custo x benefício.

Capítulo 6

Conclusão

Na atual economia da informação as organizações estão ficando cada vez mais dependentes das redes e sistemas de computadores. Eles permeiam cada aspecto da atividade humana envolvido com informação e são freqüentemente os componentes mais críticos de uma empresa e por isto mesmo merecem muita atenção.

Como as redes estão se tornando maiores e mais complexas, o projeto e gerenciamento de sistemas está se tornando uma tarefa cada vez mais desafiante. Novas aplicações e tecnologias de comunicação estão sendo constantemente exploradas tornando cada vez mais difícil para as organizações decidir qual são as ideais para o seu uso. Sendo o suporte para os seus negócios, o risco de uma falha causada por um desempenho pobre da rede, ou até mesmo uma indisponibilidade da mesma, pode ocasionar sérias repercussões.

Como alternativa para esta situação este trabalho propôs a utilização de técnicas de avaliação e planejamento de maneira a prever o comportamento futuro do ambiente. Para tanto, foi utilizada a técnica de simulação com a criação de um modelo de desempenho que permite avaliar diferentes alternativas e cenários. Desta maneira é possível analisar o comportamento do sistema sem a necessidade de construí-lo. Pode ser caro, impraticável ou até impossível construir três alternativas diferentes, escolher a melhor e descartar as demais. Alternativamente, planos de modificação de sistemas existentes podem ser realizados através de simulação sem alterar o ambiente atual.

Os recursos computacionais têm uma capacidade finita e detectar este limite, prever sua ocorrência é a essência do planejamento de capacidade. É fundamental para uma organização ter alternativas que atendam a real necessidade e com custos menores. As soluções devem atender a demanda atual e tendências de crescimento futuro.

6.1 Dificuldades encontradas

Durante a realização dos trabalhos várias dificuldades foram encontradas. Pode-se citar:

- Dificuldade na obtenção dos dados para a realização do processo de caracterização da carga de trabalho;

- Em função de limitações da ferramenta de modelagem, algumas características importantes não puderam ser consideradas;

- A limitação da ferramenta em termos de número de componentes não permitiu modelar os clientes de acesso discado.

6.2 Sugestões para trabalhos futuros

Algumas atividades de pesquisa e desenvolvimento associadas ao tema podem ser realizadas a partir dos resultados e contribuições deste trabalho, podendo-se citar:

- Realizar um planejamento de capacidade em um ambiente de *e-business*;

- Realizar um planejamento de capacidade em um ambiente de ensino à distância;

- Modelar um ambiente para avaliar as diversas políticas de substituição de arquivos avaliando os seus impactos no desempenho;

- Caracterizar a carga de trabalho pelos diferentes usuários pois os mesmos exibem diferentes comportamentos, utilizando os serviços de diversas maneiras e freqüências. Alguns utilizam os serviços de maneira intensa, enquanto outros usam apenas eventualmente. Pode-se caracterizar o comportamento do usuário em grupos, com características semelhantes e avaliar o seu impacto no ambiente.

7 - Referências Bibliográficas

ARLITT *et al* (1999) ARLITT Martin, FRIEDRICH Rich and JAIN Tai. ***Workload Characterization of a Web Proxy in a Cable Modem Environment.*** Hewlett-Packard Laboratorios. Palo Alto. CA. 1999.

ARLITT & FRIEDRICH (1998) ARLITT Martin, R. Friedrich, and T. Jin. ***Using Workload Characterization to Improve Proxy Cache Management.*** Technical. Report HPL-98-07. Hewlett-Packard Laboratories. 1998.

ARLITT & WILLIAMSON (1997a) ARLITT Martin and WILLIAMSON Carey L.. ***Internet web servers: workload characterization and performance implication*** IEEE/ACM. Transactions of Networking. vol 5 nr.5. pp 631-645. Out 1997.

ARLITT & WILLIAMSON (1997b) ARLITT, Martin F. and Carey L. Williamson. ***Trace-Driven Simulation of Document Caching Strategies for Internet Web Servers.*** Simulation vol.68. Jan. 1997. pp.23-33.

ARLITT & WILLIAMSON (1996) ARLITT Martin F. and WILLIAMSON. Carey L.. ***Web server workload characterization.*** Proceedins of the 1996 SIGMETRICS Conference on Measurement and Modeling of Computer Systems. May 1996.

BROWNING (1995) BROWNING, Tim. ***Capacity Planning for Computing System.*** Academic Press. 1995.

BUSARI (1999) BUSARI Mudashiru. ***Performance Issues in Web Proxies.*** Univeristy of Saskatchewan. Canadá. 1999.

CACI (1998) CACI Productos Company. **COMNET III** – ***The tool for network simulations results – Reference Guide.*** 1998.

COCKCROFT (1997) COCKCROFT Adrian. ***Dissecting proxy Web cache performance.*** www.sunworld.com/swol-07-1997/swol-07-perf.html.

CROVELLA & BESTAVROS (1997) CROVELLA M, BESTAVROS A. ***Self-similarity in World-Wide-Web traffic: evidence end possible cause.*** IEEE/ACM. Transactions networking. vol 5 nr. 6. pp 835-846. Dez 1997.

Centro de Informações Internet Brasil – ***Guia do Empreendedor Internet/Brasil***

CUNHA (1998) CUNHA, Jane F.. ***Avaliação de Desempenho de comutadores ATM.*** Dissertação de Mestrado. Universidade Federal de Santa Catarina. Florianópolis (SC). 1998.

DODD (1998) DODD, Annabel. ***The Essential Guide to Telecommunications.*** Prentice Hall PTR. 1998.

DOMANSKI (1999) DOMANSKI, Bernie. ***Distributed Capacity Planning.*** *Enterprise Systems* Journal. v14.pg 481. Jan 1999.

FELDMANN *et al* (1999) FELDMANN A, CACERES R, DOUGLIES F, Glass G, RABINOVICH M. ***Performance of web proxy caching in heterogeneous bandwidht environments.*** Proceedings of IEEE infocom 99. New York. NY.pp 107-116. Mar 1999.

JAIN (1991) JAIN, Raj. ***The art of computer systems performance analysis : techniques for experimental design, measurement, simulation, and modeling.*** John Wiley & Sons. Inc. USA. 1991.

LIU & MORDOWITZ (1998) LIU, Y.H.; LI, Chris; MORDOWITZ, V.; et al. ***Distributed Simulation of a Wide-Band Digital Radio Network.*** Simposium on Performance Evaluation of Computer and Telecommunication Systems. 1998.

LOPES (2000a) LOPES, Fábio. ***O avanço do cable modem.*** Network Computing. pg 8. Edição Ago 2000. www.networkcomputing.com.br

LOPES (2000b) LOPES, Fábio. ***Não saia do ar.*** Edição Mar 2000. www.networkcomputing.com.br

MAHARTI & WILLIANSON (1999) ANINBAN Maharti, WILLIANSON Carey. ***Web Proxy Workload Characterization.*** Department of Computer Science. University of Saskatchewan. Fev. 1999.

MARKATOS (1996) MARKATOS, Evangelos P. ***Main Memory caching of Web Documents.*** Computer Networks and ISDN Systems. vol.28. pp.893-905.1996.

MENASCÉ & ALMEIDA (2000) MENASCÉ Daniel A, ALMEIDA Virgílio A F. ***Scaling for e0business – Tecnologies, Models, Performance and Capacity Planning***. Prentice Hall PTR, New Jersey, USA 2000

MENASCÉ & ALMEIDA (1999) MENASCE Daniel A, ALMEIDA Virgilio A F. ***Evaluting Web-Server Capacity.*** www.webtechniques.com/archives/1999/04/menasce. 1999.

MENASCÉ & ALMEIDA (1998) MENASCÉ, Daniel A.; ALMEIDA, Virgílio F. ***Capacity Planning for Web Performance - Metrics, Models, and Methods.*** Prentice Hall PTR. New Jersey. USA. 1998

MENASCÉ & PERAINO (1999) MENASCÉ Daniel A, PERAINO, Robert, DINH Nikki, DINH Quant. ***Planning the Capacity of a Web Server: an Experience Report.*** 1999

MOKHTAR & MOUFTAH (1999) MOKHTAR, Ahmed; MOUFTAH, H.T. ***Cooperative Caching Based on ICP and CARP Interaction.*** Symposium on Performance Evaluation of Computer and Telecommunication Systems. 1999

MURTA *et al* (1998) MURTA, Cristina D., Virgílio A.F. Almeida e Wagner M. Jr.. ***Analyzing Performance of Partitioned Caches for the World Wide Web.*** Position paper at W3C Third International WWW Cache Workshop. Manchester. England. Jun 1998.

MURTA *et al* (1999) MURTA, Cristina D., Virgilio A.F. Almeida e Wagner M. Jr.. ***Efficient Storage Management in World Wide Web Caches.*** Proceedings of the ITC'16 - 16th International Teletraffic Congress. Elsevier Science. Edinburgh International Conference Center, pages 1189-1198. UK. Jun 1999.

RENAUD (1994) RENAUD, Paul E.. ***Introdução aos sistemas cliente/servidor – Guia prático para profissionais de sistemas.*** Livraria e Editora Infobook S. 1994.

SALSBURG (1997) SALSBURG Michael A. ***Importing ViewPoint Data into ViewPoint Capacity Planner.*** www.datametrics.com, 1997

SOARES *et al* (1995) SOARES, Luiz F. G. LEMOS, Guido; COLCHER, Sérgio. ***Redes de Computadores***. Editora Campus, 1995.

SLOTHOUBER (1998) SLOTHOUBER Louis P. ***Um modelo de desempenho de um servidor WEB***. StarNine Technologies, 1998

TOPKE (1999) TOPKE Claus Rugani, ***Provedor Internet - Arquitetura e Protocolos.*** MAKRON Books. São Paulo. 1999.

YU (1998) YU Linping. ***Internet Web Proxies – Workload Characterization.*** Department of Computer Science. University of Saskatchewan. Dez 1998.

ZWIEBACK ZWIEBACK Dave. ***Web Capacity Planning: How to plan for server growth.*** www.samag.com/archieve/0704/feature.sntm

ZHAO (1999)ZHAO, Yanping.. ***Trace-Driven Simulation of Caching Strategies for Internet Web Proxy.*** Department of Computer Science. University of Saskatchewan. 1999.

WALDNER (1997) WALDNER, Rudy. ***Capacity planning for web applications.*** www.ibm.com. 1997.

WILLIAMS *et al* (1996) WILLIAMS S., M.Abrams, C.R. Standbridge, G.Abdulla and E.A.Fox. ***Removal Policies in Network Caches for World-Wide Web Documents***.Proceedings of the ACM Sigcomm96. Stanford University. Aug 1996.